Diversity of Methods and Materials in Digital Human Sciences

Proceedings of the Digital Research Data and Human Sciences DRDHum Conference 2022, December 1-3, Jyväskylä, Finland

University of Jyväskylä
Faculty of Humanities and Social Sciences

DIVERSITY OF METHODS AND MATERIALS IN DIGITAL HUMAN SCIENCES

Proceedings of the Digital Research Data and Human Sciences
DRDHum Conference hosted by
University of Jyväskylä, Finland
December 1–3, 2022

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-9450-1
ISBN (PDF): 978-951-39-9450-1
URN:ISBN: 978-951-39-9450-1

Cover Design: Jarmo Harri Jantunen

Cite: Jantunen, J.H., Kalja-Voima, J., Laukkarinen, M., Puupponen, A., Salonen, M., Saresma, T., Tarvainen, J. & Ylönen, S. (Eds.) 2022. *Diversity of Methods and Materials in Digital Human Sciences. Proceedings of the Digital Research Data and Human Sciences 2022, December 1-3, Jyväskylä, Finland*. Jyväskylä: University of Jyväskylä.
http://urn.fi/URN:ISBN:978-951-39-9450-1

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

FIN-CLARIAH

Federation of Finnish Learned Societies

Organizing Committee and Editors of the Proceedings:
Jarmo Harri Jantunen, chair, editor-in-chief
Johanna Kalja-Voima
Matti Laukkarinen
Anna Puupponen
Margareta Salonen
Tuija Saresma
Jenny Tarvainen
Sabine Ylönen

Conference Secretariat:
Terhi Paakkinen
Maija Lappalainen

Technical Editors of the Proceedings:
Minna Hautaniemi
Sanna Nordlund

Scientific Committee:
Jonne Arjoranta, University of Jyväskylä
Kati Dlaske, University of Jyväskylä
Benedikt Ehinger, University of Stuttgart
Antti Gronow, University of Helsinki
Lauri Haapanen, University of Jyväskylä
Niko Hatakka, University of Helsinki
Annika Herrmann, University of Hamburg
Tuomo Hiippala, University of Helsinki
Ari Huhta, University of Jyväskylä
Pasi Ihalainen, University of Jyväskylä
Tommi Jantunen, University of Jyväskylä
Tommi Jauhiainen, University of Helsinki
Antti Kanner, University of Helsinki
Kristiina Korjonen-Kusipuro, University of Jyväskylä
Timo Korkiakangas, University of Helsinki
Raine Koskimaa, University of Jyväskylä
Liisa Kääntä, University of Vaasa
Salla-Maaria Laaksonen, University of Helsinki
Veronika Laippala, University of Turku
Esa Lehtinen, University of Jyväskylä
Mietta Lennes, University of Helsinki
Krister Lindén, University of Helsinki
Janne Matikainen, University of Helsinki
Katja Mäntylä, University of Jyväskylä
Matti Pohjonen, University of Helsinki
Juho Pääkkönen, University of Helsinki
Päivi Rainò, Humak University of Applied Sciences
Anna Rantasila, University of Tampere
Maria Ruotsalainen, University of Jyväskylä

Elina Salomaa, University of Jyväskylä
Daniel Shanahan, Northwestern University
Laura-Elena Sibinescu, University of Helsinki
Valtteri Skantsi, University of Oulu
Minttu Tikka, Aalto University
Tuukka Ylä-Anttila, University of Helsinki
Klaus Zechner, Educational Testing Service, Princeton
Elaine Zosa, University of Helsinki

TABLE OF CONTENTS

PART II METHODS AND TOOLS

Multimodal, multilingual and mixed methods approaches to communication

Developing practical tools and data

Preface

Even though some disciplines have been working with digital data, resources, and technology for decades already, there is today a growing number of ways in which the human and social sciences use them to investigate different aspects of human life. Researchers now produce and have access to an increasing variety of digital materials which reach across new fields, offering new opportunities for interdisciplinary scholarly work and enabling certain data to be approached from different viewpoints and with novel and diverse methodologies.

The Conference for *Digital Research Data and the Human Sciences* (DRDHum 2022) brought together researchers with different areas of interest and expertise to discuss the themes of data compilation and management, and to share their knowledge and experience of using digital materials, methods, and data analysis tools. The conference was held in Jyväskylä, Finland, on 1st – 3rd December, 2022. The event, originally planned for the year before but postponed due to the COVID-19 pandemic, was the second in a series of conferences hosted every two years by a different university in the FIN-CLARIAH Consortium. The first (D)RDHum Conference was held in 2019 at the University of Oulu, with a focus specifically on linguistic text corpora. In DRDHum 2022, the scope of the presentations went beyond textual data and corpus analysis, to also explore, e.g., online multimodal video data and mixed-method approaches.

This year's conference was successful and very international: there were altogether 65 oral presentations in sections and in four workshops, 10 poster presentations, and a total of 151 participants coming from over twenty different countries. The texts in this volume cover the various repercussions of using and interacting with digital information and amply show the diversity of participants, both in terms of their disciplines and approaches. As digitalization takes on more roles, it is clear we could be entering a new era of big data and dataism; making it all the more crucial that the humanities and social sciences continue to study digital phenomena in a multidisciplinary manner.

All participants whose submissions were accepted and presented at the conference were also given the opportunity to author an article to be published in the conference proceedings, and eighteen of these were picked for this volume. Each published paper went through a double-blind peer-review process, assessed by two members of the DRDHum Scientific Committee. We take this opportunity to thank all participants for their contributions to the conference, to the authors for the articles published here, and the members of the Scientific Committee for reviewing the papers and providing valuable recommendations.

This volume follows a structure aligned with the thematic focus of each article. Part I consists of contributions in which thematically different aspects of our social reality are examined using digital data and various digital research methods. This section starts with discussions on political communication and polarization. First, the timely article by *Ascone*, *Placzynta* and *Troschke* presents a mixed-methods corpus approach to study online antisemitic speech. This is followed by *Reveilhac* and *Schneider's* study, in which Kernel density estimation and topic modelling is used to examine how politicians debate policy issues in parliamentary sessions and on Twitter. Topic

modelling is also employed in *Kahle* and *Kliche's* analysis of migration- and refugee-related interviews with organization representatives. An algorithm analysis, in turn, is used by *Chuang* to show how political polarization and conflicts of interest emerge on YouTube, leading into a discussion of the role of digital media in political communication.

The second section in Part 1 looks more precisely at online communication, equality and well-being. A corpus-analysis of simulated physician-patient interactions, especially their pragmatic elements, e.g., code-switching and deixis is presented by *Mány et al.*, while *Galkin* uses digital ethnography to analyse the importance of online health communities for older people – especially the implications of such communities in remoter rural areas. Rural/urban issues are also touched with regard to equality in *Korjonen-Kuusipuro, Tuuva-Hongisto* and *Berg's* study on the day-to-day digital life of young people. Finally, Part I concludes with two articles thematically linked by fashion. *Sánchez Casanova* conducts a nethnographic analysis of secondhand clothing entrepreneurs in the virtual space of Instagram, while *Li* discusses the role of blogs as virtual spaces for tattoo wearers to express themselves.

Part II looks at the methods and tools developed for analysing digital materials and online communication. The first thematic section is on multimodal, multilingual and mixed method approaches to online communication. It starts with *Gonzalez's* discussion on methods and tools used in the creation of online web apps that can be applied in speech data and sociophonetic research. The article by *Hutin* and *Allassonnière-Tang* studies how crowd-sourced digital data from Lingua Libre can be used to investigate acoustic measurements; while sign-language videos provide the data for a discussion by *Hernández et al.* that combines linguistics and neuroscience to investigate sign language processing in the brain. Meanwhile, the next article, by *Jackson*, discusses whether multimodal actor-network theory and methods are suitable for analysing online role-playing games. Finally, *Roschier* proposes steps towards a methodology that would combine both qualitative and quantitative approaches in the study of ancient Latin texts.

The next thematic section is on issues related to new data and tools being developed in the digital human sciences. An article by *Salonen et al*. relates the process of building a multimedia Finland-Swedish Sign Language Corpus and how, in doing so, the language of a deaf community can be preserved and revitalized. In the next article, *Moreau*, *Vogel* and *Walsh* look at the increasingly relevant ethical issue of how data from social media is compiled and analysed in the context of research into attitudes towards technology among older adults. The last two articles comprise the final thematic section on language learning from the perspective of assessment and academic writing. *Von Zansen* and *Huhta* illustrate the process of building an automated tool for assessing the oral skills of language learners, while *Pace-Sigge* and *Sumakul* use keyword and cluster analyses to compare a novel textbook produced using artificial intelligence algorithms with traditionally produced language textbooks.

This collection of articles barely scratches the surface of the wealth of digital materials and methods presently employed in the human sciences. Nevertheless, we hope that the breadth of expertise and innovative results contained in this collection will, lead to new data collections, help further develop creative new methodological practices, and suggest new research questions in this broad field that studies different aspects of social and cultural human life.

On behalf of the editorial board,
Jarmo Harri Jantunen, Matti Laukkarinen, Tuija Saresma and Jenny Tarvainen

Antisemitic Hate Speech in Social Media Comments: An Interdisciplinary Mixed-Methods Corpus Study

Laura Ascone, Karolina Placzynta, Hagen Troschke[1]

Zentrum für Antisemitismusforschung (ZfA), Technische Universität Berlin

E-mail: laura.ascone@decoding-antisemitism.eu

Abstract

Online antisemitism, among other forms of hate speech, has been increasingly visible; investigating its emergence and spread is therefore crucial. The pilot project "Decoding Antisemitism: An AI-driven Study on Hate Speech and Imagery Online" aims to analyse the conceptual characteristics, frequency, and linguistic structure of online antisemitism, with the eventual objective of using machine learning to develop an algorithm capable of recognising its explicit and implicit forms. Since the project is still in progress, this contribution focuses on its first, qualitative stage, the insights gained in the process, and its value for the forthcoming steps of the project. It also showcases the multi-step methodological design, which seeks to capture the complexity of antisemitic speech online.

The initial focus has been on text analysis of user comments found on the websites and social media platforms of mainstream media in the UK, Germany and France, triggered by the media coverage of domestic and international events. The comments are collected with a custom-designed data crawling tool; so far, 77,000 of the collected comments have been qualitatively analysed using MAXQDA software and following detailed coding guidelines to categorise every antisemitic utterance in the dataset. The guidelines currently comprise over 150 precisely defined categories, encompassing both conceptual and linguistic structures.

Our analyses to date have confirmed the necessity of an interdisciplinary qualitative approach: on the one hand, this allows to describe and characterise the appearance of antisemitism in specific discourses; on the other, it sheds light on aspects of verbal antisemitism that might have gone unnoticed if only a quantitative approach had been adopted. The size of our annotated dataset, combined with the prominent share of qualitatively analysed implicit speech within it, make the data unique and particularly valuable for the training of AI models.

Keywords: antisemitism, anti-semitism, qualitative content analysis, online hate speech, social media, corpus studies

1. Research Background and Aims

Since their emergence on the web, social networks have elicited diverging reactions and opinions. While they are appreciated for helping create new forms of social relations, they are also criticised for facilitating both the emergence and circulation of hate speech. It is acknowledged that one of the main causes is the possibility to remain anonymous on the web (Mondal, Silva & Benevenuto 2017), which may lead internet users to convey hate ideas that they would not have dared to express in offline contexts. Even though in many cases web users' names sound real, we cannot say with certainty who they are; this effectively hinders research into the intended meaning behind their individual actions. Moreover, on social networks, users are likely to join communities (e.g., *Facebook* groups) whose members share the same interests[2]. As a consequence, the interaction with people sharing the same point of view tends to strengthen both the link among the users and the ideas they adhere to. This mutual reinforcement is particularly dangerous when internet communities build themselves around a hate ideology. Furthermore, the spread of hate speech has been accelerating also in mainstream online environments (Monnier & Seoane 2019), sometimes as a consequence of being brought into the discourse by public figures who are regarded as authorities. Here, in order to avoid being affected by moderation, users may be led to resort to linguistic and discursive strategies such as puns and allusions. This way, their hate messages can go unnoticed by the moderators.

[1] Authors listed in alphabetical order.

[2] For echo chambers and filter bubbles, see Pariser (2011) and Cinelli et al. (2021).

Among other forms of hate speech, online antisemitic content has been increasingly visible; investigating its emergence and spread is therefore crucial. Schwarz-Friesel and Reinharz defined verbal antisemitism as follows:

> *All linguistic utterances that devalue, stigmatize, discriminate against, and defame Jews qua Jews can be considered forms of verbal antisemitism. […] Verbal antisemitism can thus be defined as consisting of all formulations that explicitly and/or implicitly express stereotypes about Jews, share anti-Jewish conceptualizations and feelings, and pass along traditional Judeophobic images of the world. (Schwarz-Friesel & Reinharz 2017, 19)*

Even though this definition captures key elements of verbal antisemitism, it is important to highlight the fact that antisemitic discourse can target non-Jewish actors as well. Among them, we find verbal attacks targeting players working against antisemitism or supporting Israel, as well as verbal forms of secondary antisemitism (see, among others, Rensmann 2004; Bergmann 2010; Beyer 2015). Online antisemitism has been analysed qualitatively and/or quantitatively (and sometimes combined with machine learning) by ADL (2018), Allington (2018), ADL (2019), Barna & Knap (2019), CST (2019), Jikeli, Cavar & Miehling (2019), Schwarz-Friesel (2019), Allington & Joshi (2020), Ozalp et al. (2020), Zannettou et al. (2020), Becker (2021), Becker, Ascone & Troschke (2022), Becker & Troschke (2022a), Jikeli et al. (2022), Riedl et al. (2022), and an overview of antisemitism in several social media milieus is given in Hübscher & von Mering (2022).

Our pilot research project "Decoding Antisemitism: An AI-driven Study on Hate Speech and Imagery Online" has paid particular attention to verbal antisemitism; the qualitative analysis presented here forms part of our ongoing work. In it, we aimed to analyse the conceptual characteristics, frequency, and linguistic structure of online antisemitism. Rather than the ideological reasons behind its recent rise, or individual web users' intentions, the analysis focused on patterns of expression and their potential interpretations and impact on readers. The analyses were conducted on user comments found on the websites and social media platforms of mainstream media in the UK, Germany and France, adopting a mixed-methods approach and using a detailed guidebook designed by the research team to annotate and analyse verbal antisemitism; in this paper, we also offer our insights into our interpretation processes and methods of validation. Content analyses (Mayring 2015) of both antisemitic topoi and linguistic phenomena will eventually be combined with quantitative analyses, which focus on frequency, collocation and vector analyses, with the eventual objective of using machine learning to develop algorithms capable of recognising explicit and implicit forms of antisemitic speech online.

2. Research Design

The "Decoding Antisemitism" project follows a mixed-methods approach with a three-step process that consists of qualitative and quantitative analyses as well as machine learning[3]. While all three will be briefly presented below to give an overview and context of our research, the qualitative analysis is the step that we will discuss here in detail, since the machine learning stage has only recently begun and most quantitative analyses are still in the future at this point.

[3] Various aspects of our study design and qualitative analyses also appear in Becker, Troschke & Allington (2021), and Becker & Troschke (2022b).

2.1. Qualitative Analysis

For qualitative content analyses, the project examines current antisemitic topoi and accompanying patterns of language use in web user comments. Apart from the qualitative step being a necessity for gathering data for machine learning, it serves the pursuit of independent analytical goals. We want to find out which discourses trigger what kind and amount of antisemitic reactions, how the utterances through which they are communicated are built linguistically, how antisemitic topoi appear together, how they are distributed across the platforms, and much more[4].

2.2. Machine Learning

In the second step, the results of these qualitative analyses are used for machine learning. The extensive and context-sensitive categorisation of web comments in the qualitative step provides the basis for using these categorised data as training input in the development of algorithms able to carry out the task of distinguishing between antisemitic and non-antisemitic comments with increasing accuracy. To this end, several already available classifiers – types of machine learning algorithms – will be tested, both on their own and in combination, for their capabilities to identify antisemitism in texts. Due to the complexity of language use and the associated challenges for an algorithm to classify texts correctly, this step will be continuously supervised and include numerous test runs. The tests are first carried out on our uncoded raw data, from which the samples of the qualitatively analysed data are taken. If the classifiers perform well, the tests are extended to text data from social media not used in our qualitative research. In initial tests with different classifiers based on 20,000 coded English-language comments, logistic regression performed best in identifying antisemitic comments with an F1-value of 0.752 (Jansen 2022). In the course of the research training, data will be continuously updated with categorised user reactions on new discourse events in order to reflect possible changes in the antisemitic repertoire and broaden the range of utterances the classifiers learn to recognise.

2.3. Quantitative Analysis

As a third step, alongside qualitative examinations the team also carries out regular quantitative analyses on these coded data, e.g., for correlations between specific antisemitic topoi, parallels and contrasts between individual discourse events, news outlets and countries with regard to these topoi. Using all the raw data collected throughout the project, further quantitative analyses will be conducted based on R, Python, and Lancsbox. This will allow us to build statistical models tracing patterns in antisemitic communication on corpus level, such as triggers for antisemitic comments, key concepts and words, collocations, n-grams, and more.

We have decided on this methodical arrangement, moving from qualitative to quantitative examination, due to the repeated observation that antisemitic utterances overwhelmingly appear in implicit forms in segments of the political mainstream (see, among others, Schwarz-Friesel & Reinharz 2017; Becker 2021; Giesel 2021)[5]. An antisemitic topos can be verbalised in a great variety of forms – most of which differ to a great extent from explicit or extreme utterances. This means that

[4] For examples of such qualitative analyses see Becker, Ascone & Troschke (2022), Becker & Troschke (2022b) and our biannual discourse reports (https://decoding-antisemitism.eu/publications/#discourse-reports).

[5] There are three possible causes for the communication of antisemitic attributions in an implicit or coded way: it can serve to protect the reputation or the self-image of the writer (if this is seen as a necessity in certain milieus) and it can help to avoid sanctions from moderators; it follows general aspects of language economy (preferring implicit forms for the sake of efficiency); conveying information in coded form can increase its effectiveness for the recipient (since, in this way, writers allow readers to participate in "secret knowledge").

when searching antisemitic utterances based on certain words (profanities, terms connected to open reproduction of antisemitic stereotypes, or explicit death wishes), a large proportion of the comments in which antisemitism is communicated cannot be found and therefore is not taken into account. Qualitative analyses demonstrate that comments sections where a predefined search for predetermined deductive categories, words, or word combinations resulted in few or no hits can nevertheless contain large numbers of antisemitic statements. They also illustrate that the constitution of antisemitic attributions can be semantically so open that conspicuous word accumulations – or even relevant terms – can be completely absent.

3. Data Collection

3.1. Selecting Sources and Data Samples

In order to select material for analysis, our team[6] continuously monitored the news in the mainstream media in France, Germany and the UK. We then chose specific events for in-depth analysis, based not just on how broadly they were covered by the media, but also, crucially, on the reactions they triggered from web users in comments sections; more specifically, on the potential prominence of antisemitic speech in their responses. Such events can be both country-specific, such as the court proceedings against the former concentration camp personnel in Germany in 2021, or the heated debate in France about the introduction of the Covid-19 "passport" the same year, and of international importance – for example, the escalation phase of the Arab-Israeli conflict in May of 2021, which was widely reported in French, German and British media at the time.

We obtained our data from media outlets positioned in the traditional political mainstream, with around twelve such outlets per country (including *The Daily Mail*, *The Guardian*, *The Independent*, *The Telegraph* in the UK; *Le Monde*, *Libération*, *Le Figaro*, *Le Parisien*, in France, and *Bild*, *Spiegel*, *Süddeutsche Zeitung*, *Welt* in Germany, to name a few). We also collected data from their official social media accounts, primarily from *Facebook* and *Twitter*, since some media outlets outsource the comment function to their social media pages, as is the case with *The Guardian*, *Libération*, or the *Süddeutsche Zeitung*, or place it behind the paywall. The comments we took into consideration were those posted under articles relating to our selected discourse event.

The choice of the traditional political mainstream as our data source inevitably meant that the data we collected contained lower levels of antisemitism and fewer extreme examples of it, compared to more radical sources. Additionally, the content we analysed had already been moderated by either human moderators (as in the case of *Twitter*) or automatic ones (e.g., on *Facebook*). Nevertheless, we found that the comments sections within this milieu contained a significant amount of antisemitic hate speech, often implicit and therefore more challenging to identify. Our aim was precisely to shed light on antisemitic hate speech in conventional online spaces, especially when it evades detection because of its coded forms and its normalisation.

3.2. Building a Corpus

In the next step, the comments were collected with a custom-designed data crawling tool. Comment order in the downloaded threads differs according to the way the respective sources structure them

[6] The project team comprises nine researchers from a range of academic backgrounds, including cognitive linguistics, sociolinguistics, social semiotics, pragmatics, (critical) discourse analysis, multimodality, corpus linguistics, media studies, as well as critical theory, history, language and iconography of antisemitism, racism, political extremism, and populism. The team is divided into three country teams, working in close contact with one another.

online (newest, most approval, etc.). This order is kept and guarantees comparability within threads from one source (e.g., *Facebook*). That way, we are creating a vast library of files in four text formats. Each of the files represents a thread of comments under a news article, and can be imported into the content analysis software MAXQDA (see section 4.1. Guidebook and Analysis Software). In order to build a robust, well-balanced dataset for a given discourse event, we selected threads that represented a cross-section of the mainstream in terms of political alignment, and – if possible – a mix of social media comment threads and news website threads. We aimed to analyse the same number of comments from each thread, and use the first comments appearing in each text file in order to avoid bias. In most discourse events, we analysed several thousand comments per one discourse event on average, with our largest discourse event corpus to date comprising 15,000 comments; particular attention was paid to this specific event, which deals with the escalation phase in the Arab-Israeli conflict (May 2021), because of its high potential in eliciting antisemitic reactions. This allowed us to collect a large number of antisemitic comments that will be required by and used in the machine learning process[7]. So far, the entire project corpus comprises 77,000 online comments qualitatively analysed and annotated by our interdisciplinary team of expert researchers.

3.3. Example of a Discourse Event Corpus

In May 2021, the eleven-day escalation period of the Arab-Israeli conflict generated a significant amount of media coverage in German, French and British mainstream press; we also observed a great deal of debate in comments sections. In order to maintain comparability between the three national discourses, we sought to build a corpus using only threads from *Facebook* pages of leading media outlets in the three countries, posted under articles covering the same events, as a great deal of the comment activity seemed to focus on that platform in all three countries. The resulting corpus comprised a total of 4,520 comments, just over 1,500 per each country. In two cases, 150 first comments were collected from each of ten threads, in one case – 100 first comments from each of fifteen threads, as the available threads were slightly shorter on average. Next to first level comments, these also included reactive second level and third level comments, downloaded by our crawling tool in the same order each time.

By using the same social media platform as our source, the same discourse trigger, and the same number and similar distribution of comments, we built a balanced dataset of a considerable size. After downloading the threads in the form of text files, and following a qualitative examination of the comment sample, we were able to compare and contrast the three sub-corpora in terms of the proportion of antisemitic comments, and frequency of specific antisemitic topoi (Becker et al. 2021).

4. Qualitative Analysis Process

4.1. Guidebook and Analysis Software

We follow Mayring's approach to qualitative content analysis (supported by a pragmalinguistic focus) and apply it to our corpora[8]. Qualitative content analysis has to be based on a set of rules that guarantee the consistency of data interpretation. To achieve this, a guidebook with annotating – or

[7] The proportionally large share of this topic in the total coded data reflects the fact that Israel-related antisemitism is the most widespread form of antisemitism and must be included in the training data for the classifiers to a corresponding extent.
[8] For methodological literature on qualitative content analysis see, for example, Mayring (2015), Krippendorff (2004) and Kuckartz (2018).

coding – instructions is needed, containing the categories according to which the data are examined. These categories are provided with definitions and illustrative anchor examples in order to clearly distinguish them from each other, and to minimise deviations in interpretation, and thereby categorisation, as much as possible. The guidebook consists of several sections which can be – or have to be – applied simultaneously to a certain comment: machine learning codes, general classification of the comment (antisemitic, counter speech or simply non-antisemitic), topoi (antisemitic and non-antisemitic), linguistic (such as implicitness, speech acts, word choice) and semiotic categories (emoticons, GIFs, memes and other kinds of images, as well as typographic characters). The guidebook is constructed using both deductive and inductive categories. The former includes the special codes for machine learning and definitions of categories that were set in advance according to what we expected to find in the data with higher probability, e.g., specific antisemitic topoi, forms of implicitness. The inductive categories are developed differently: every time we encounter a phenomenon (topos, linguistic or semiotic aspect) that is relevant for our research questions but is not yet in the guidebook, we gather incidences of that phenomenon in our data and craft a (preliminary) definition for this phenomenon, usually revised at a later point when we have gained more experience coding it, which includes possible recoding of the data coded so far with this phenomenon. This openness to inductive categories gives the opportunity not only to make new or unexpected phenomena visible, but also to include the entire spectrum and latest dynamics of antisemitic speech into our machine learning data for the classifiers to operate with the most comprehensive data possible. Both types of categories are based on existing research literature (see e.g., Schoeps & Schlör 1996; Julius 2010), with the exception of some inductive categories for which we had to craft definitions entirely based on our findings.

In the MAXQDA software, our categories are arranged in a code system representing our guidebook. In most cases, coding a comment combines multiple codings from the several sections. Code co-occurrences illustrate how certain topoi are communicated linguistically within an utterance and how certain topoi appear combined. The various extraction possibilities of conceptual-linguistic/semiotic combinations or combinations of topoi support us in drawing conclusions from the coded data.

4.2. Identifying Antisemitism

Our operationalisation of antisemitism comprised antisemitic topoi (stereotypes, conspiracy theories, forms of Nazi analogy, etc.), insults – whether specifically antisemitic in themselves, or aimed at Jewish identity – and various speech acts that express the wish to harm Jews or Israelis for their Jewish or Israeli identity. To illustrate our approach to coding, we present one of the rules according to which we distinguish between antisemitic and non-antisemitic attributions: if any of the above is used explicitly against Jewish individuals or groups, it was coded as antisemitic. Even if a particular writer unintentionally expressed themselves in this way, it is highly possible that many readers of the comment would associate the negative attribution with the person or group's Jewishness. For example, "Soros is the evil of the world/the most evil person" is an antisemitic utterance, because it reflects the stereotype which alleges that Jews are malevolent by nature and responsible for various global calamities[9]. If, on the other hand, a comment criticised an individual or a group, their practices or professional actions, however harsh, it was not regarded as antisemitic, even if the underlying worldview of the web user – which we cannot infer directly from the comment – may be antisemitic.

[9] For more examples of antisemitic utterances in comments and their interpretation see Becker et al. (2021), Becker, Troschke & Allington (2021), Ascone et al. (2022), Becker, Ascone & Troschke (2022), and Becker & Troschke (2022b).

"Soros is an evil banker" is an example of a strongly negative evaluation that had to be interpreted conservatively as an attribution of his actions as an investment banker, since it is explicitly linked to his professional background. Moreover, negative evaluations towards this particular profession are quite typical. With our cautious approach, we would categorise the latter as non-antisemitic – even though the combination of the concepts JEW, MONEY/CAPITALISM, and EVIL[10] reaches into the centre of antisemitic imagery and could well be perceived – and intended – as such.

4.3. Interpretation Process

We extrapolated the meaning of utterances based on the combination of knowledge regarding (conventionalised forms of) language, as well as contextual and world knowledge. Language knowledge was the most basic asset required – preferably on a native speaker level. Contextual knowledge comprised the content of the article or post that initiated the comment thread, and the content of other relevant comments in the thread. Both content resources can be indispensable to complete the meaning of a comment when it refers to them. What we considered as world knowledge is all content knowledge not derived from a contextual relation within a comment thread: it consists of information about topoi, their (semantic) relations, scripts, processes, and schemes in which the cultural knowledge about their environment is organised[11]. It serves, for example, to situate topoi, e.g., within the realm of antisemitism using our expert knowledge related to this, or to "complete" implicatures[12], that is to enrich such an utterance with the hinted to information.

Language and world knowledge were relevant to every interpretation. It is impossible to identify antisemitism without world knowledge. Certainly, an explicit attribution can be understood from a linguistic point of view, and without being augmented by world knowledge – however, the utterance still needs to be situated within the ideology of antisemitism. Contextual knowledge is very often a relevant source for interpretation, but there are statements that can be comprehensively interpreted without contextual information because their meaning does not rely on, or relate to, the context and would be the same in other contexts.

In order to extrapolate a comment's meaning, we broke it down into units of meaning, determined which propositions were present, tracked how it was embedded in the context, and identified its linguistic forms, as well as any informational gaps that had to be filled by conclusions. We then summarised the conclusions for the individual components of the comment, and how they related to each other. This entire process typically took place internally but could be supported with notes.

The application of the knowledge from these three areas determined the extent to which all meanings and nuances are identified when categorising comments. Insufficient knowledge as well as incomplete or faulty reasoning processes on the linguistic and conceptual level can lead to certain meanings not being recognised, or being projected onto a comment without sufficient evidence. However, the effect of over-interpretation can also be the result of an over-sensitivity caused by priming: continuous examination of the subject matter in the coding process can lead to false presumptions of the (not reliably verifiable) presence of the subject in the comment. Distortions are possible in two directions: either overlooking antisemitic meanings, or false positives.

That is why it was crucial that statements were categorised conservatively. This meant deciding

[10] Since stereotypes are phenomena that exist on a conceptual, i.e. mental, level and can be reproduced using language, stereotypes are given in small caps on the following pages in accordance with the conventions of cognitive linguistics.
[11] For more details see Schwarz-Friesel (2013, 37–41).
[12] For the characteristics of implicatures see Levinson (1983).

in favour of the more likely reading (or the reading with less deduction steps necessary for obtaining it) in situations where there were at least two possible interpretations of an ambiguous utterance, in order to prevent false positives. Conflicting valid interpretations were set out, the probabilities of the correctness of different interpretations weighed against each other, and the use of world knowledge to enrich each interpretation was undertaken with caution, as adding a number of deductive steps to back an interpretation could, when combined, reduce its likelihood. The guidebook supported this process with its clearly defined interpretation scheme, which allowed us to arrive at meaningful interpretations, and also helped to avoid mis- or over-interpreting the statement (false positives).

Our validation experiences have shown that when the same text corpus was categorised by several coders looking for certain content, contextual knowledge and the respective levels of topic-related world knowledge were the most likely to produce differences in our conclusions. The extent of world knowledge has a decisive impact on the comparability of reasoning and coding processes between coders, and the resulting categorisations of texts. Therefore, the level of this world knowledge has to be raised collectively in advance. However, the more evenly our world knowledge is distributed among us, the more contextual knowledge remains the issue of differing interpretations, which in turn has a decisive effect on the interpretation of a comment. That means a faulty interpretation of a comment can generate consequential errors – sometimes even resulting in a cascade effect of wrong interpretations of comments building on each other.

4.4. Validation

In order to ensure the comparability and quality of the interpretational approach of the different coders, we have used two types of validation. In the initial test phase of coding, all the researchers from the three country teams coded the same comments sections. At first, the material was restricted to English – the language shared by all of us, and the main working language of the project. This step was necessary for introducing all the coders to the guidebook in the same way, and to establish shared interpretations and a coding workflow based on the guidebook instructions. It also helped us to gather feedback on the guidebook, as a result expanding and improving its definitions. Later on, validation involved each of the project's three working languages. In the first months of the coding process, we exclusively used consensual validation (Bortz & Döring 2006), the process of reaching intersubjective agreement between coders via discussion of the categorisations of texts. In the test period, it took place in a succession of short intervals. Once the regular coding phase started, intervals became longer but included double-checking all codings of one coder by another. Controversial cases were discussed within teams of three, or – if not solved there or of special interest – collectively by the whole research team.

Once a reliable common understanding had been established after the first six months of the project, we scaled back consensual validation and started to determine the coding quality using the calculation of intercoder reliability for each team and discourse event[13]. We opted for the percentage agreement approach and set our threshold for an acceptable overall intercoder agreement across all codes that are chosen for the test at 80%[14]. This percentage is a minimum – and a compromise. On the one hand, we need the data to be as "clean" as possible – also for the machine learning step of the project. On the other, the complexity of utterances we seek to cover by a vast number of codes necessarily leads to diverging interpretations of certain subtleties. In case a pair of coders missed the 80% agreement in such a test, they had to rely again on consensual validation (with double-checking

[13] We used R for collating intercoder agreement results (Vincent 2022).
[14] A discussion of intercoder reliability and its approaches can be found in O'Connor & Joffe (2020).

all codings) of the corpus the test relates to.

We also regularly collated intercoder reliability for the whole research team, preventing the country teams from developing diverging understandings of codes in the course of their collective work and exchange. This ensured that while the whole team never coded one dataset together, all coded data had the necessary consistency to be part of the same machine learning process. In the whole team intercoder reliability tests, we considered only the general classification of comments. However, the detailed levels of topoi and linguistics are touched upon when we discuss those of the comments which produced disagreement. Discussions following both types of validation processes may show that guidebook instructions need to be adjusted. Therefore, validation is also valuable in this respect.

5. Conclusions

The project "Decoding Antisemitism: An AI-driven Study on Hate Speech and Imagery Online" aims to analyse the verbal and visual expressions of antisemitism in user comments posted on the websites and social media platforms of mainstream media in the UK, Germany and France. We have outlined here the broader mixed-methods approach of our text-related research and presented the way we collected our raw data, and then analysed it qualitatively during the first stage of our research project. On the one hand, the qualitative approach allowed us to raise awareness of the diversity and complexity of antisemitism on the level of single utterances, as well as of the characteristics of antisemitic communication in the context of particular web discourse. On the other, it will allow us to produce a large amount of qualitatively coded data in order to train algorithms to automatically recognise antisemitic comments based not on keywords, but rather on the entirety of word material obtained in the interpretation process operated by our team of researchers, and will provide input for the planned quantitative analyses.

Furthermore, the same approach can be adapted to the study of other hate ideologies that have recently been increasing on the web, providing tools and insights for identifying both their specificities and similarities. In this way, the project can contribute to future interdisciplinary research beyond antisemitism studies.

6. Acknowledgements

The pilot research project "Decoding Antisemitism: An AI-driven Study on Hate Speech and Imagery Online" is funded by the Alfred Landecker Foundation.

References

ADL (2018). *Quantifying Hate: A Year of Anti-Semitism on Twitter*. https://www.adl.org/resources/reports/quantifying-hate-a-year-of-anti-semitism-on-twitter

ADL (2019). *Gab and 8chan: Home to Terrorist Plots Hiding in Plain Sight*. https://www.adl.org/resources/reports/gab-and-8chan-home-to-terrorist-plots-hiding-in-plain-sight

Allington, D. (2018). 'Hitler had a valid argument against some Jews': Repertoires for the Denial of Antisemitism in Facebook Discussion of a Survey of Attitudes to Jews and Israel. – *Discourse, Context & Media* 24, 129–136. https://doi.org/10.1016/j.dcm.2018.03.004

Allington, D. & Joshi, T. (2020). 'What others dare not say': An Antisemitic Conspiracy Fantasy and Its YouTube Audience. *Journal of Contemporary Antisemitism* 3 (1), 35–54. https://doi.org/10.26613/jca/3.1.42

Ascone, L., Becker, M. J., Bolton, M., Chapelan, A., Krasni, J., Placzynta, K., Scheiber, M., Troschke, H., & Vincent, C. (2022). *Decoding Antisemitism: An AI-driven Study on Hate Speech and Imagery Online.*

Discourse Report 3. Berlin: Technische Universität Berlin. Center for Research on Antisemitism. https://decoding-antisemitism.eu/publications/third-discourse-report/

Barna, I. & Knap, Á. (2019). Antisemitism in Contemporary Hungary: Exploring Topics of Antisemitism in the Far-Right Media Using Natural Language Processing. *Theo-Web* 18 (1), 75–92. https://doi.org/10.23770/tw0087

Becker, M. J. (2021). Antisemitism in reader comments: Analogies for reckoning with the past. Cham: Palgrave Macmillan.

Becker, M. J., Allington, D., Ascone, L., Bolton, M., Chapelan, A., Krasni, J., Placzynta, K., Scheiber, M., Troschke, H., & Vincent, C. (2021). *Decoding Antisemitism: An AI-driven Study on Hate Speech and Imagery Online*. Discourse Report 2. Berlin: Technische Universität Berlin. Center for Research on Antisemitism. https://decoding-antisemitism.eu/publications/second-discourse-report/

Becker, M. J., Ascone, L. & Troschke, H. (2022). Antisemitic comments on Facebook pages of leading British, French, and German media outlets. *Humanities and Social Sciences Communications* 9 (339). https://doi.org/10.1057/s41599-022-01337-8

Becker, M. J. & Troschke, H. (2022a). How Users of British Media Websites Make a Bogeyman of George Soros. *Journal of Contemporary Antisemitism* 5 (1), 49–67. https://doi.org/10.26613/jca/5.1.100

Becker, M. J. & Troschke, H. (2022b). Decoding Implicit Hate Speech – Using the Example of Antisemitism. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (eds.), *Challenges and perspectives of hate speech analysis: An interdisciplinary anthology.* Berlin: Digital Communication Research. https://doi.org/10.48541/dcr.v12.0 [In print]

Becker, M. J., Troschke, H. & Allington, D. (2021). *Decoding Antisemitism: An AI-driven Study on Hate Speech and Imagery Online.* First Discourse Report. Berlin: Technische Universität Berlin. Center for Research on Antisemitism. https://decoding-antisemitism.eu/publications/first-discourse-report/

Bergmann, W. (2010). Sekundärer Antisemitismus. [Secondary Antisemitism] In W. Benz (ed.), *Handbuch des Antisemitismus. Bd. 3, Begriffe, Theorien, Ideologien* [Handbook of Antisemitism. Vol. 3, Terms, Theories, Ideologies]. Berlin, Boston: De Gruyter Saur, 300–302.

Beyer, H. (2015). Theorien des Antisemitismus: Eine Systematisierung. [Theories of antisemitism. A systematisation]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* (67), 573–589. https://doi.org/10.1007/s11577-015-0332-7

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* [Research methods and evaluation for human and social scientists]. Heidelberg: Springer.

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118 (9). https://www.pnas.org/doi/full/10.1073/pnas.2023301118

CST (2019). *Engine of Hate: The Online Networks behind the Labour Party's Antisemitism Crisis*. https://cst.org.uk/news/blog/2019/08/04/engine-of-hate-the-online-networks-behind-the-labour-partys-antisemitism-crisis

Giesel, L. (2021). Comparisons between Israel and Nazi Germany in Contemporary German Discourse. In A. Lange, K. Mayerhofer, D. Porat, & L. H. Schiffman (eds.), *An End to Antisemitism!, vol 3: Comprehending Antisemitism through the Ages: A Historical Perspective.* Berlin, Boston: De Gruyter, 443–464. https://doi.org/10.1515/9783110671995

Hübscher, M., & von Mering, S. (eds.) (2022). *Antisemitism on Social Media*. Abingdon, New York: Routledge.

Jansen, F. (2022). *KI-gestützte Antisemitismus-Erkennung* [AI-based detection of antisemitism], unpublished bachelor's thesis. Hochschule für Technik und Wirtschaft (HTW), Berlin, Fachbereich 4: Informatik, Kommunikation und Wirtschaft, Studiengang Angewandte Informatik.

Jikeli, G., Axelrod, D., Fischer, R. K., Forouzesh, E., Jeong, W., Miehling, D., & Soemer, K. (2022). Differences between antisemitic and non-antisemitic English language tweets. *Computational and Mathematical Organization Theory*, 1–35. https://doi.org/10.1007/s10588-022-09363-2

Jikeli, G., Cavar, D. & Miehling, D. (2019). *Annotating Antisemitic Online Content. Towards an Applicable Definition of Antisemitism*. https://arxiv.org/abs/1910.01214

Julius, A. (2010). *Trials of the Diaspora. A History of Anti-Semitism in England*. Oxford: Oxford University Press.

Krippendorff, K. (2004). *Content Analysis. An Introduction to its Methodology*. 2nd ed. Thousand Oaks: Sage.

Kuckartz, U. (2018). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung (Grundlagentexte Methoden)* [Qualitative content analysis. Methods, practice, computer support (Basic texts on methods)]. 4th ed. Weinheim, Basel: Beltz.

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Mayring, P. (2015). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* [Qualitative content analysis. Basics and techniques]. 12th ed. Weinheim, Basel: Beltz.

Mondal, M., Silva, L. A., & Benevenuto, F. (2017). A measurement study of hate speech in social media. *Proceedings of the 28th ACM conference on hypertext and social media*, 85–94. https://doi.org/10.1145/3078714.3078723

Monnier, A. & Seoane, A. (2019). Discours de haine sur l'internet [Hate speech on the Internet]. *Publictionnaire. Dictionnaire encyclopédique et critique des publics*. http://publictionnaire.huma-num.fr/notice/discours-de-haine-sur-linternet/

O'Connor, C. & Joffe, H. (2020). Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. *International Journal of Qualitative Methods* 19, 1–13. https://doi.org/10.1177/1609406919899220

Ozalp, O., Williams, M. L., Burnap, P., Liu, H. & Mostafa, M. (2020). Antisemitism on Twitter: Collective Efficacy and the Role of Community Organisations in Challenging Online Hate Speech. *Social Media + Society* (April/June), 1–20. https://orca.cf.ac.uk/132742/1/2056305120916850.pdf

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. London: Penguin.

Rensmann, L. (2004). *Demokratie und Judenbild. Antisemitismus in der politischen Kultur der Bundesrepublik Deutschland* [Democracy and the image of Jews. Antisemitism in the political culture of the Federal Republic of Germany]. Wiesbaden: VS Verlag für Sozialwissenschaften.

Riedl, M. J., Joseff, K., Soorholtz, S. & Woolley, S. (2022). Platformed antisemitism on Twitter: Anti-Jewish rhetoric in political discourse surrounding the 2018 US midterm election. *New Media & Society* 1 (21). https://doi.org/10.1177/14614448221082122

Schoeps, J. H. & Schlör, J. (eds.) (1996). *Antisemitismus. Vorurteile und Mythen* [Antisemitism. Prejudices and myths]. München: Piper.

Schwarz-Friesel, M. (2013). *Sprache und Emotion*. [Language and Emotion]. 2nd ed. Tübingen, Basel: Francke.

Schwarz-Friesel, M. (2019). 'Antisemitism 2.0.' The Spreading of Jew-Hatred on the World Wide Web. In A. Lange, K. Mayerhofer, D. Porat & L. H. Schiffman (eds.), *An End to Antisemitism!, vol. 1: Comprehending and Confronting Antisemitism*. Berlin, Boston: De Gruyter, 311–338. https://doi.org/10.1515/9783110618594-026

Schwarz-Friesel, M. & Reinharz, J. (2017). *Inside the Antisemitic mind: The language of Jew-hatred in contemporary Germany*. Waltham: Brandeis University Press.

Vincent, C. (2022). *Intercoder Agreement [Source code]*. https://github.com/ChloeVincent/IntercoderAgreement

Zannettou, S., Finkelstein, J., Bradlyn, B. & Blackburn, J. (2020). A Quantitative Approach to Understanding Online Antisemitism. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (1), 786–797. https://ojs.aaai.org/index.php/ICWSM/article/view/7343

Complementing Kernel Density Estimation and Topic Modelling to Visualise Political Discourse

Maud Reveilhac[1] & Gerold Schneider[2]

[1] Lausanne University, Institute of Social Sciences; [2] Zurich University, Department of Computational Linguistics

E-mail: maud.reveilhac@unil.ch

Abstract

We examine how politicians shape and convey policy issues among different communication channels and how well we can visualise issue ownership and issue framing using data-driven methods. Drawing from a political communication approach, we propose to complement two methods – Kernel density estimation and topic modelling – to visualise political discourse. Our case study is established on two main data sources: transcripts of parliamentary debates and tweets of Swiss elected politicians. We propose a two-step methodology. First, we use topic modelling to identify the main policy issues emphasised by politicians. Second, we use the topical content and political affiliation as meta-information in Kernel density estimation to calculate and display the distances between important features, the main extracted topic for each document, and political affiliation. We compare the obtained results from the transcripts of parliamentary debates and from Twitter data. Using conceptual visualisation maps enables us to qualitatively discuss whether the findings are indicative of issue ownership and stylistic markers pointing to issue framing. The proposed methodology can be applied to a variety of research questions in the realm of communication studies with an interdisciplinary and exploratory focus. It also provides a suitable way to account for polarisation processes that can be linked to affordances of communication channels and to institutional specificities of the national political context.

Keywords: topic modelling, distributional semantics, data visualisation, interdisciplinarity, qualitative interpretation

1. Introduction

In this paper, we propose to combine two data-driven methods that enable us to make sense of a large amount of written natural language in the realm of political communication. First, we rely on Topic Modelling (TM), specifically using the Latent Dirichlet Algorithm (LDA), to extract policy issues from tweets emitted by politicians. Second, we rely on visualisations based on Kernel Density Estimation (KDE) while using the extracted topics and politicians' party affiliation as meta-information to produce conceptual maps.

To the best of our knowledge, the combination of TM and conceptual maps with KDE is new. This combination can complement existing techniques to visualise the results from TM (e.g. tethne and pyLDAvis) and previous studies showing the utility of adding meta-information to conceptual maps for more qualitative interpretation (Taavitsainen et al. 2019; Schneider 2020; Schneider 2022; Eve 2022).

The combination of TM and KDE allows us to observe issue ownership and to explore issue framing. In a nutshell, while issue ownership indicates *what* politicians are most likely to talk about, issue framing tells us *how* they talk about policy issues. To investigate both political communication patterns, conceptual maps are produced relying on the visualisation tool textplot (McClure, 2015), which converts a document into a network of terms, thus displaying the high-level topic structure of the text in the spirit of Digital Humanities and Moretti's distant reading (Moretti 2013).

We address the following research questions: RQ1) Can we reliably visualise issue ownership and issue framing by using KDE with meta-information about main topical content and political affiliation? RQ2) To what extent is the proposed visualisation methodology robust across data sources? To answer these RQs, we combine TM and KDE. While TM assumes context beyond the document (e.g. the same topic is witnessed by many documents), KDE can possibly add situational knowledge. Indeed, the observation window plays a central role in KDE, whereas it is ignored when conducting

TM. We thus propose to use the information of TM (contextual variable) and of party affiliation (individual variable) as meta-information in KDE. Conceptual maps then enable us to visualise more precisely issue ownership and issue framing with semantic patterns. To evaluate the quality and validity of the visual outputs, we need to interpret the found patterns carefully and further validate our interpretation, notably by relying on statistical tests.

Our study has important implications. From a methodological point of view, it is important to reflect on each data source's potential for the study of political communication as digital humanities increasingly relies on large datasets of textual data (Tyrkkö 2020), notably to study political discourse (Curran, Higham & Ortiz2018). We also propose a new way to visually display political communication by combining two data-driven methods that are used in digital humanities. While TM assign words to topics and thus partly solves the issue of mapping words to concepts, the distances in the distributional space between words and topics are not visible and the ordering of the documents according to relevant dimensions (e.g. topic content, party affiliation, time, etc.) is ignored. These semantic distances are best displayed in maps of concepts using KDE. We discuss how our methodology enables to model the theories of issue ownership and issue framing.

2. Related Work

2.1. Opportunities for Data-driven Methods in Digital Humanities

With data-driven approaches, such as TM and KDE, patterns typically emerge from the analysed data. Such approaches have the advantage that they can find new patterns or confirm old ones while partly shifting the process of hypothesis generation. As emphasised by Grimmer, Roberts and Stewart (2022, 14–15), the availability of textual data for conducting social and political research encourages researchers to adopt a more sequential approach between a theory-driven deductive approach and a more data-driven inductive approach. In our study, we propose a sequence of tasks: 1) we summarize the textual meaning using TM; 2) we use the results from TM, as well as politicians' party affiliation, as meta-information to conduct KDE analysis; 3) we produce conceptual maps to investigate issue ownership and issue framing; 4) we validate the content of the maps using statistical tests.

We suggest that the outputs from TM can profitably feed KDE to gain a more fine-grained level of insight. This is in line with the perspective proposed by Evans (2022, 9) stating that "[t]he power of many of these methods comes from their ability to plug-and-play in recursive designs crafted to the research task". Indeed, TM provides the best possible content summary of a large collection of documents, but it says little about how a given topic can be framed differently according to individual or contextual factors. Using the topic weights enables us to investigate the prevalence of topics according to some factors of interest. However, it is less suitable to assess how a common topic (e.g. environment) is framed by different factors (e.g. party affiliation). To do so, we need a more transparent visualisation of the content of single topics.

Social sciences and digital humanities are increasingly turning into subdisciplines of data science and statistics, where data-driven approaches are increasingly used. Data-driven approaches can deliver trends of word features where data space and interpretations would otherwise be challenging. Therefore, visualisation techniques for displaying many of the salient features are important and there is a need to discuss the validity of the obtained maps.

To date, evaluations of results from TM exist and exploit the measurement of semantic similarity (Röder, Both & Hinneburg 2015). However, TM provides us with little information about how the topics are related to each other. KDE can tackle this issue by uncovering associations and

relations between topics, but it is even harder to evaluate the quality of KDE output automatically. We propose to take advantage of the strengths of each method to bring different aspects of the framing of political discourse to the surface. We should be cautious that these methods are also sensitive to pre-processing decisions (e.g. the size of the documents has a large influence). Furthermore, the random component of the LDA of TM needs to be assessed critically. Moreover, KDE performs reasonably well even with small amounts of data. KDE maps can be interpreted globally as similar topics are placed close to each other, the opposing ends of the map often corresponding to thematic opposites.

2.2. Opportunities Associated with Each Data Sources

To date, there are several national databases enabling work with digital parliamentary data. For instance, infrastructures exist to enable the access to parliamentary datasets[15]. All interfaces provide meta-information – such as party affiliation, name of the speaker, and some other contextual information –, while some also include data-analytic tools in addition to search and data exploration facilities. Furthermore, other databases have enriched parliamentary data with external data sources, such as the project Linked Data of the European Parliament (Van Aggelen et al. 2017). Furthermore, common schemes have been developed to annotate the databases in a shared and structured manner, such as the Parla-CLARIN initiative[16]. A detailed description of these and other databases can be found in the proceedings of the workshop Digital Parliamentary Data in Action (La Mela, Norén & Hyvönen 2022), which presents results from ongoing research on creating and using historical and present parliamentary data. The studies cover a wide range of research interests within the study of parliamentary culture and politics, as well as that of the use of language.

Transcripts of parliamentary debates may suffer from several inconveniences when studying political communication. First, the communication protocols prevailing for organising the parliamentary sessions and the political culture may strongly impact the prevalence of the discussed policy issues and also the way politicians express their opinion during the debates. For example, a political culture aiming to come to a compromise and within a multi-partisan setting may favour consensual discourse and a balanced time repartition between the different speakers or partisan groups. In this regard, Switzerland represents an interesting case to study, notably due to the distinctive consensus-oriented character of its political system (Arend 1999), which enables Swiss politicians to have very low access barriers to the public debate. Second, the textual material may be heavily impacted by transcription rules, thus removing stigmatized, unconventional, and impolite language (Mollin 2007).

With respect to social media data, the Twitter Parliamentarian Database[17] is a multi-source and manually validated database of parliamentarians on Twitter across 26 countries. It aims to promote comparative and transnational analysis. In this view, the database also incorporates additional information from other expert databases[18].

While it is the most comprehensive database about politicians' social media behaviour and

[15] Examples include: Canadian Linked Parliamentary Data Project (https://lipad.ca/), the portal of the Italian parliament (https://storia.camera.it/), the German Bundestag documents (https://opendiscourse.de/), as well as the UK Hansard open user interface (https://github.com/stephbuon/hansard-shiny).
[16] https://www.clarin.eu/resource-families/parliamentary-corpora
[17] http://twitterpoliticians.org
[18] This includes: the Manifesto Project Database on political parties (https://manifesto-project.wzb.eu/), the Electoral System Design Database on the countries' electoral and legislative systems (https://www.idea.int/data-tools/data/electoral-system-design), the Chapel Hill Expert Survey on party positions on specific issues (https://www.chesdata.eu/), and the ParlGov database on political parties, elections and cabinets (https://www.parlgov.org/).

discourse, Twitter is arguably only one possible social media choice for politicians. Although Twitter is much used by "elite" users, such as politicians and journalists, there are several limitations worth considering when using this data source for political analysis. For instance, politicians who are active on Twitter may possess very different characteristics than politicians who are not using Twitter, or not using social media at all. Furthermore, politicians have very different levels of involvement on Twitter. Moreover, the collection of tweets in a corpus is dependent on the API for sampling the tweets (e.g. streaming API versus historical API). However, tweets may also disappear as they are removed by the authors or cancelled by the platform. Finally, metadata about politicians is usually scarce on Twitter as the default profile, location, and description fields are not necessarily filled by politicians. Therefore, a lot of manual work is needed to get the necessary meta-information either before collection the data (e.g. list of politicians with Twitter accounts) or after (e.g. political leaning or other political views).

2.3. The Usefulness of Data-driven Methods to Identify Policy Framing

Text-as-data approaches have increasingly been applied to evaluate research questions revolving around issue framing, agenda setting, or issue definition (Grimmer & Stewart 2013; Gilardi & Wüest 2018). For instance, the Policy Frames Project relies on machine learning to track media tone and framing of policy issues in news (Card et al. 2015). In our study, we propose to investigate issue ownership and issue framing by using conceptual maps, while adding contextual information from topic modelling and individual information from partisan affiliation.

Intuitively, we know that politicians from specific parties tend to dominantly speak about certain issues (e.g. Green Parties focus on climate and environment issues, while left-leaning politicians highlight social and labour market policies, and right-leaning politicians emphasise concerns related to immigration and asylum). However, aside from their preferred "owned" party agenda (Bélanger & Meguid 2008), politicians also express views on policy issues that are not typically associated with their party program. This leads us to the idea of policy framing. Policy framing can be defined as "the act of emphasizing certain aspects or dimensions of an issue" (Sides 2006, 426) to influence the public perception of a given policy issue. Framing can be conceived as a refinement of issue ownership to the extent that politicians can selectively "highligh[t] some feature of the issue on which [they are] likely to be regarded as more competent" (Petrocik 1996, 829). The use of frames is a well-known method from media studies (Entman, Matthes & Pellicano 2009) and is defined as generic schemata of interpretation (Goffman 1974).

It is important to consider that textual data are characterised by a mix of public and personalised communication, and blurred boundaries between political expertise and entertainment, all of which affect the status of a hegemonic data source in comparative research of political communication (Enli & Skogerbø 2013). In parliamentary debate, politicians must discuss policies and find compromises to propose solutions. On social media, politicians are also prompted to engage in shared discussions by reacting to events or polemics, by interacting with other (opponent) users, or by defending a different point of view on common themes of discussion. From a substantive point of view, linking politicians' emphasis on issues with their political affiliation and the type of communication channel used (e.g. parliamentary debate and social media discussions) has important consequences for the transmission of citizens' interests and the legitimation of authority in policy-making (Søyland 2022). For example, the way in which politicians formulate their statements in parliamentary debates may be more informative about the partisan framing of important policy issues discussed by the entire partisan spectrum, while social media posts may provide more useful information about how

politicians impose their partisan agenda in a protocol-free online environment (Castanho Silva & Proksch 2021).

Social media data have been used to study the extent to which politicians are responsive to the priorities of the public (Barberá et al. 2019), its role in political agenda setting (Gilardi et al. 2022), as well as partisan framing of political debate (Stier 2016). Social media is also useful for the study of populist attitudes in political communication. For example, studies have recently started to examine how populist communication logic and online opportunities might go hand in hand (Engesser et al. 2017; Ernst et al. 2017; Jacobs & Spierings 2019; Mazzoleni & Bracciale 2018; Schwarzbözl & Fatke 2017) by analysing the extent to which online opportunity structures can foster populist communication and the diffusion of populist ideas to broader public audiences.

3. Data Collection and Methods of Analysis

3.1. Data Sources

With respect to transcripts of parliamentary debates, we rely on the Official Bulletin of the Swiss Parliament made available after every parliamentary session at the National Council and at the Council of the States[19]. The transcripts include every elected politician's speech including meta-information such as the canton and the party affiliation of the speaker. The data we use cover the period from December 2011 to December 2019. To extract the data, we downloaded all pdf files available and used the R language to extract the party members' contributions using regular expressions. We extracted the text, name of the speaker, and political affiliation. The corpus was thus composed of 58,048 members' contributions stemming from the main Swiss parties: Swiss People's Party (SVP), Social Democratic Party (SP), the Liberals (FDP), Conservative Democratic Party (BDP), Christian Democratic People's Party (CVP), Evangelical People's Party (EVP), Green Party (GPS) and the Green Liberal Party (GLP). We grouped the Green parties (GPS and GLP) together under the label Gruene and group the centrist parties (CVP and EVP) under the label Center.

Concerning Twitter data, the first step in the data gathering process is defining which accounts need to be included. For the sake of comparability with the transcripts of parliamentary debates, we included the Twitter accounts of every national elected politician in the sample of Twitter users. To collect the data, we retrieved all the available tweets relying on the R package *academictwitteR*[20]. We filtered only the tweets emitted by politicians when they were active in Parliament. Our final corpus contained 298,280 unique tweets (including original tweets, retweets, and replies).

In both corpora (parliamentary data and Twitter data), German utterances represented about 70% of the content in the three main Swiss languages (German, French and Italian). On Twitter, the repartition was 69% German, 19% French, and 12% other. In parliamentary debates, the repartition was 72% German, 25% French, and 3% other. We selected all data in German and French and we used GoogleTranslate to translate the French utterances into German. To minimize translation issues, we conducted several pre-processing steps (especially relevant for Twitter data) including: the removal of URLs and the separation of concatenated words (e.g. *ClimateChange* becomes *climate change*). Before conducting KDE, we conducted additional pre-processing steps, including lemmanisation, the replacement of accentuated characters (e.g. *ö* becomes *oe*), the removal of punctuation, as well as the converting of all words into lowercase letters. Because tweets are very short, we concatenated unigrams with bigrams to add contextual information. We finally added the

[19] https://www.parlament.ch/fr/ratsbetrieb/amtliches-bulletin
[20] https://cran.r-project.org/web/packages/academictwitteR/academictwitteR.pdf

political affiliation and the main topic of the document in front of every contribution to use them as a meta-information to be plotted into the semantic network.

Table 1 summarises the number of politicians with Twitter accounts and the number of politicians who contributed to the debates during the sessions of the parliamentary chambers for each partisan affiliation. It shows that politicians from the right (SVP) are less represented on Twitter compared to their share of seats in Parliament, in contrast to leftist (Gruene and SP) politicians.

Party	Parliamentarians on Twitter	Parliamentarians
SVP	32 (19%)	103 (28%)
FDP	27 (16%)	67 (18%)
BDP	6 (4%)	11 (3%)
CVP	36 (21%)	61 (17%)
EVP	2 (1%)	3 (1%)
SP	51 (30%)	86 (23%)
Greens	17 (10%)	36 (10%)
Total	171	367

Table 1: Repartition of unique parliamentarians on Twitter and number of unique parliamentarians included in our corpora according to partisan affiliation.

3.2. Data-driven Methods

3.2.1. Topic Modelling

We rely on TM to account for the extent to which a politician's emphasis on specific policy issues depends on their choice of communication channels and their political affiliation. Intuitively, the logic of TM implies that specific topics are much more likely to generate specific words (Blei 2012). The insights into the relationship between words, documents, and contexts combine document classification with the strong semantic unity of the discourse of a topic and of a document. Documents (e.g., parliamentarians' contributions and tweets) and words are given, and topics are fitted iteratively starting from a random configuration. The fitting process makes sure that the overall probability of the given documents and words is as high as possible. The user only must set the number of topics that the algorithm should use. To conduct TM, we used the LDA algorithm as implemented in the textmineR package (Jones, Doane & Jones 2021) from the R programming language. To choose the best number of topics for each data source, we ran several models iteratively to look for the best topic coherence. The best number of topics was 65 for parliamentary data and 42 for Twitter data.

Supervised document classification approaches (e.g. machine learning) have the advantage that they are easy to evaluate, however target classes (e.g. policy issues) are not necessarily clear-cut (e.g. possible overlaps within discourse) and some classes are sometimes too sparse to provide a space of semantically useful classes. TM can offer a solution to both problems as it combines keyword detection and document classification. It thus exploits contextual and semantic information beyond the level of the document to the topic level. Church (2000) observed that the likelihood of a word to appear in a document radically changes if the same word has appeared before. In TM, documents and words are given, and topics are fitted iteratively starting from a random configuration.

Once the topics had been extracted, we assessed the semantic validity of the topics manually by scanning the keywords and judging how coherent they were. Furthermore, we assessed the

predictive validity by investigating how well variation in topic usage corresponded with expected factors (e.g. party affiliation and year). After this manual verification, we classified each topic into fewer policy issue categories as we were ultimately interested in investigating the framing of policy issues by party affiliation and data sources. In order to make the distribution of topics easier to interpret, we manually classified each topic into the categories of "the most important problem" survey question used in the Comparative Candidates Survey[21]: Agriculture; Economy; Education & culture; Environment & energy; European integration; Finances & taxes; Gender issues & discrimination; Immigration & asylum; International relations & conflicts; Labour market; Law & order; Political system; Public health; Public services & infrastructure; Regions & national cohesion; Social security & welfare state. The reduction of the number of topics into a set of pregiven policy issue categories had several advantages. First, it enabled us to filter out non-substantive policy topics (e.g. topics about party campaigning activities or topics about candidates' interviews did not provide meaningful information about policy issues and were thus filtered out from further semantic analyses). Second, it enabled us to assess if human annotators could easily give a different label to each topic.

3.2.2. Kernel Density Estimation

A limitation with TM results suggests that the extracted topics tend to be broad, and the reported top keywords can sometimes be overlapping (e.g. same words can appear in different contexts and thus topics) or can appear loosely related. There is thus a need to detect semantically related words and to disambiguate word senses (Schütze 1998). In this view, the hypothesis that "[y]ou shall know a word by its company" (Firth 1957) holds the key to detecting semantically similar words. Sahlgren (2006) observes that small observation windows deliver collocations, while expanding the window allows for the inclusion of more context (e.g. synonyms, antonyms, associations).

This logic is exploited by distributional semantics (Baroni & Lenci 2010) and is very useful for the analysis of (social) media content and (political) discourse. Distributional semantics opens many analytical possibilities. Indeed, with large window sizes, the influence of the individual document and of the topic grows larger. In KDE, the window sizes are so large that they typically stretch across many documents. This has the advantage that similarity of adjacent documents can be reflected. Thus, if the documents in a collection are sorted by meaningful categories (in our case, topical content, and party affiliation), we can profit from context beyond documents. Consequently, it is possible to interpret the entire semantic space.

Like TM, KDE is also a distributional semantic method (McClure 2015; Schneider 2020; Eve 2022). Mathematically, it is a smoothing method which allows one to separate trends from noise even in sparse data situations (Zucchini 2003). It considers words that appear particularly often in a same large text segment as related. Adding the political affiliation and the main topic of the document as meta-information also allows us to account for the prevalence of issue ownership and to assess the presence of stylistic markers pointing to issue framing.

3.2.3. Conceptual maps

We generated two types of conceptual maps:

To test for issue ownership, we ordered the document collection according to the key "party-topic" in descending order of the parliamentarian's name. The variable *party* was always ordered from left-leaning to right-leaning affiliation. We investigated whether the extra structuring elements

[21] http://www.comparativecandidates.org/

(topic content and party affiliation) display patterns of issue ownership. Practically, we relied on the Python library *textplot*[22] (McClure, 2015) to calculate word-word associations. To generate the conceptual maps, we applied the spring-attraction algorithm *ForceAtlas2* (Jacomy et al. 2014) in *Gephi*[23] which displays maps of the most frequent 2,000 terms based on the KDE results with a bandwidth of 1,000 for tweets and of 2,000 for parliamentary speeches. Because we filtered out documents that were not assigned to any relevant policy issue categories during TM, we ended up with 138,117 tweets (47% of the available tweets at the TM stage) and with 31,339 parliamentary speeches (54% of the available speeches at the TM stage).

To test for issue framing, we selected tweets and speeches that had been labelled under the "EU and Europe" category by TM (12,296 tweets and 1,864 speeches) and we added the party affiliation as meta-information. We displayed only the 500 top terms in the conceptual maps.

There are a few rules for interpreting conceptual maps. Elements and word features that appear far from the centre have extreme peaks in their distribution. Furthermore, the vicinity of terms mirrors semantic relationships. Elements that appear close to the centre tend to frequently appear in all sections of the corpus. The orientation of the figure plays no role in the semantic interpretation.

4. Results

4.1. Semantic Network Testing for Issue Ownership

Figures 1 and 2 display the conceptual maps obtained from parliamentary speeches (Figure 1) and tweets (Figure 2) ordered according to the key "party-topic".

Figure 1 demonstrates that the topic prevalence and the party affiliation played only a minor role in the shaping of the conceptual map of parliamentary speeches. For instance, there is not even a clear separation between left-leaning (SP and Greens) and centre (BDP, CVP, GLP) or right-leaning (FDP and SVP) parties. Instead, most of the parties appear in a dedicated section of the map (on the right) – only Greens and BDP are drawn towards environmental issues, an SP towards gender issues and social security. However, we can observe a grouping of topics with shared issues. For instance, "immigration and asylum" appear next to the topic "EU and Europe". Similarly, "social security" and "public health" are next to each other. We can zoom-in to assess the extent to which these topic groups share a similar vocabulary. For instance, immigration and European questions are linked by words such *personenfreizugigkeit* (freedom of movement) and *entwicklungshilfe* (development aid). The fact that the shape of the conceptual map is quite round shows that there was little individualisation of the debate, and only few distinct strong clusters. Overall, processes like party ownership almost entirely dissolve in a parliamentary debate, although some topics might be more sensitive to policy framing than others.

[22] https://github.com/davidmcclure/textplot
[23] https://www.gephi.org

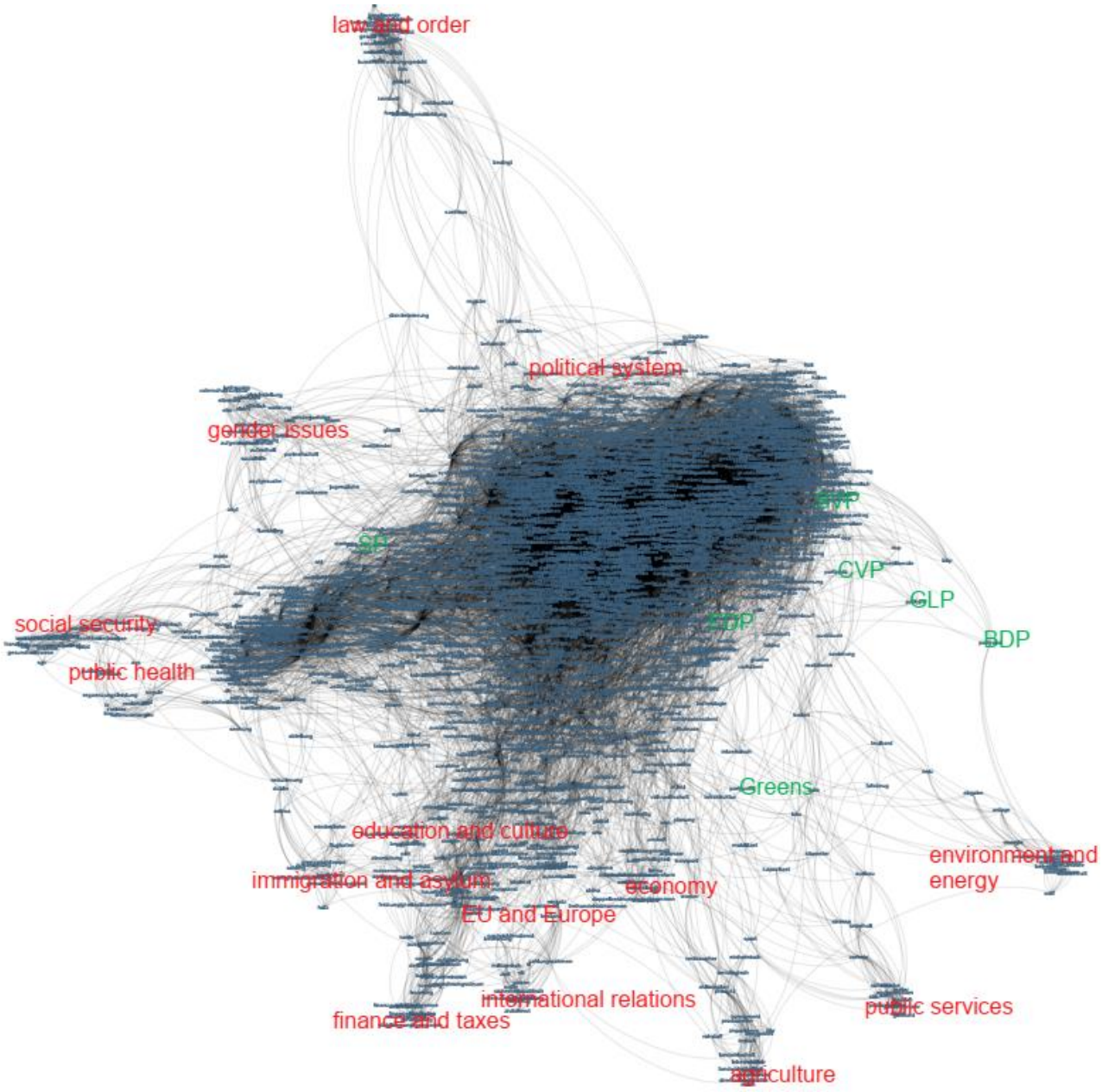


Figure 1: Semantic network from the transcripts of parliamentary debates ordered according to the key "party-topic".

Figure 2 displays the conceptual maps obtained from tweets ordered according to the key "party-topic". It demonstrates that topic prevalence and party affiliation had more influence on these semantic relations than they had in the maps based on parliamentary speeches (Figure 1). It shows which parties were likely to address policy issues, for instance: "gender issues" and "economy" were more likely to be discussed by SP, the "environment and energy" by the Greens, "public services" and "social security" by centrist parties, and "immigration and asylum" and "EU and Europe" by SVP. The fact that this figure is less round mirrors a strong semantic divide along the main axis, here left versus right, with SP and Green on the left and SVP on the right of the map.

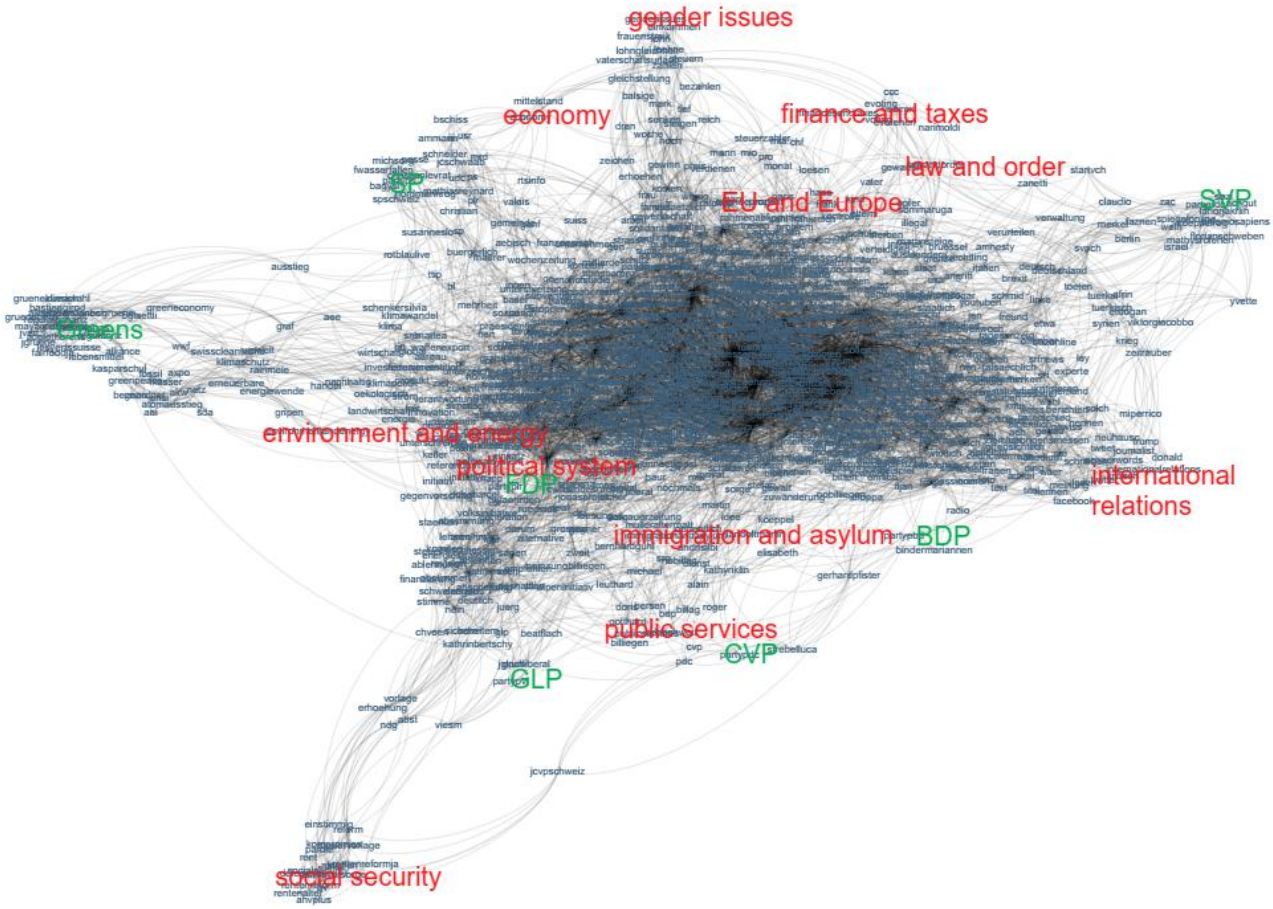


Figure 2: Semantic network from tweets from parliamentarians ordered according to the key "topic-party".

The results from the conceptual maps related to issue ownership can be assessed for their statistical significance. Using hypothesis testing, we can check whether the semantic networks are reliable in comparison to the trends that would be found in the data before they are translated into conceptual maps. Table 2 displays the statistical significance comparing proportions of topics between the SVP and the other parties. For instance, we used the two-proportions z-test for testing whether some topics are more prevalent (or "greater") in the SVP communication compared to these topics' prevalence in the communication of the other parties. A p-value less than the significance level alpha = 0.05 would conclude that the prevalence of the topics is significantly higher for the SVP. Table 2 shows that the prevalence of increased focus on "law and order" and "international relations" from the SVP compared to the other parties mirrors the findings of the conceptual maps from Figure 2. However, Table 2 shows a statistically significant prevalence for "immigration and asylum" and "international relations" in the SVP communication in Parliament which is not mirrored in the Figure 1. This could be explained by the fact that Figure 1 is still heavily influence by formal expressions (such as greetings, references the colleagues and commissions, words related to procedures, etc).

	Twitter			Parliament		
	SVP	**Other parties**	**p-value**	**SVP**	**Other parties**	**p-value**
EU & Europe	8%	9%		6%	6%	
Immigration & asylum	4%	5%		9%	5%	< 0.001
Law & order	15%	10%	< 0.001	6%	8%	
International relations	18%	12%	< 0.001	7%	5%	< 0.001

Table 2: Proportion test assessing the statistical prevalence of a sample of topics for the SVP compared to the other parties.

4.2. Semantic Network Testing for Issue Framing in Twitter Discussions

To dig deeper into issue framing, we can select tweets related to a given issue, such as "EU and Europe" to investigate its framing by different parties. Figures 3 and 4 provide the results from the parliamentary speeches (Figure 3) and from the tweets (Figure 4).

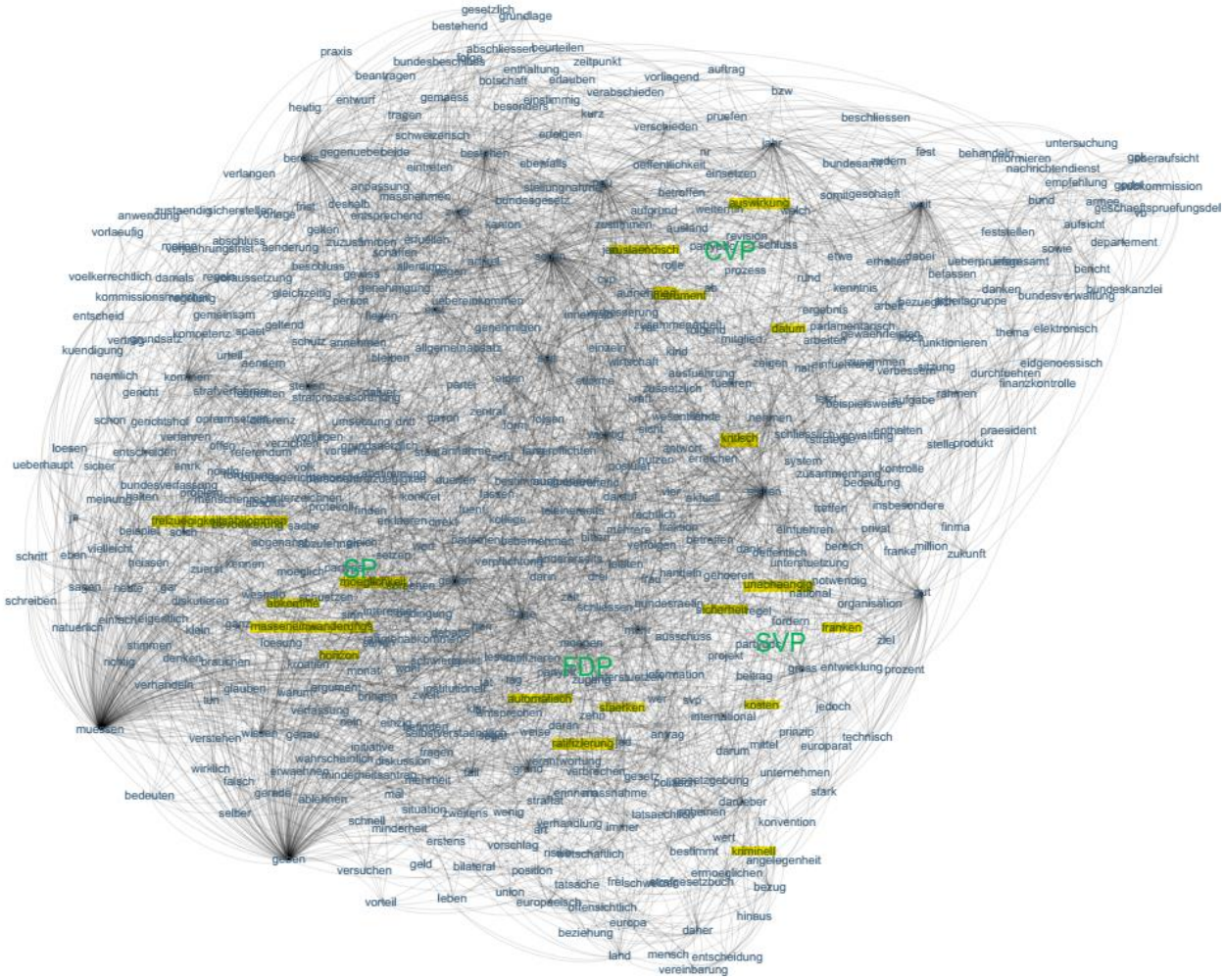


Figure 3: Semantic network from the EU-related tweets from parliamentarians.

Figure 3 demonstrates that, although we still observe a round shape like Figure 1, there are differences in partisan framing of this specific policy issue. For instance, SP used positive language about the EU (e.g., *möglichkeit* for 'possibility') and concentrated on questions related to immigration (e.g., *masseneinwanderung* 'mass migration'). The SVP used a more negative language emphasising security (e.g., *sicherheit*) and independence (e.g., *unhabhängig*), and it also underlined cost-related

issues (*kosten*). We can also note that the words referring to criminality (e.g., *kriminell*) are close to the SVP. The other parties that made it in the top terms (FDP and CVP) were generally more concerned about contractual debates and the concrete implementation of policies (*ratifizierung*, *automatisch*).

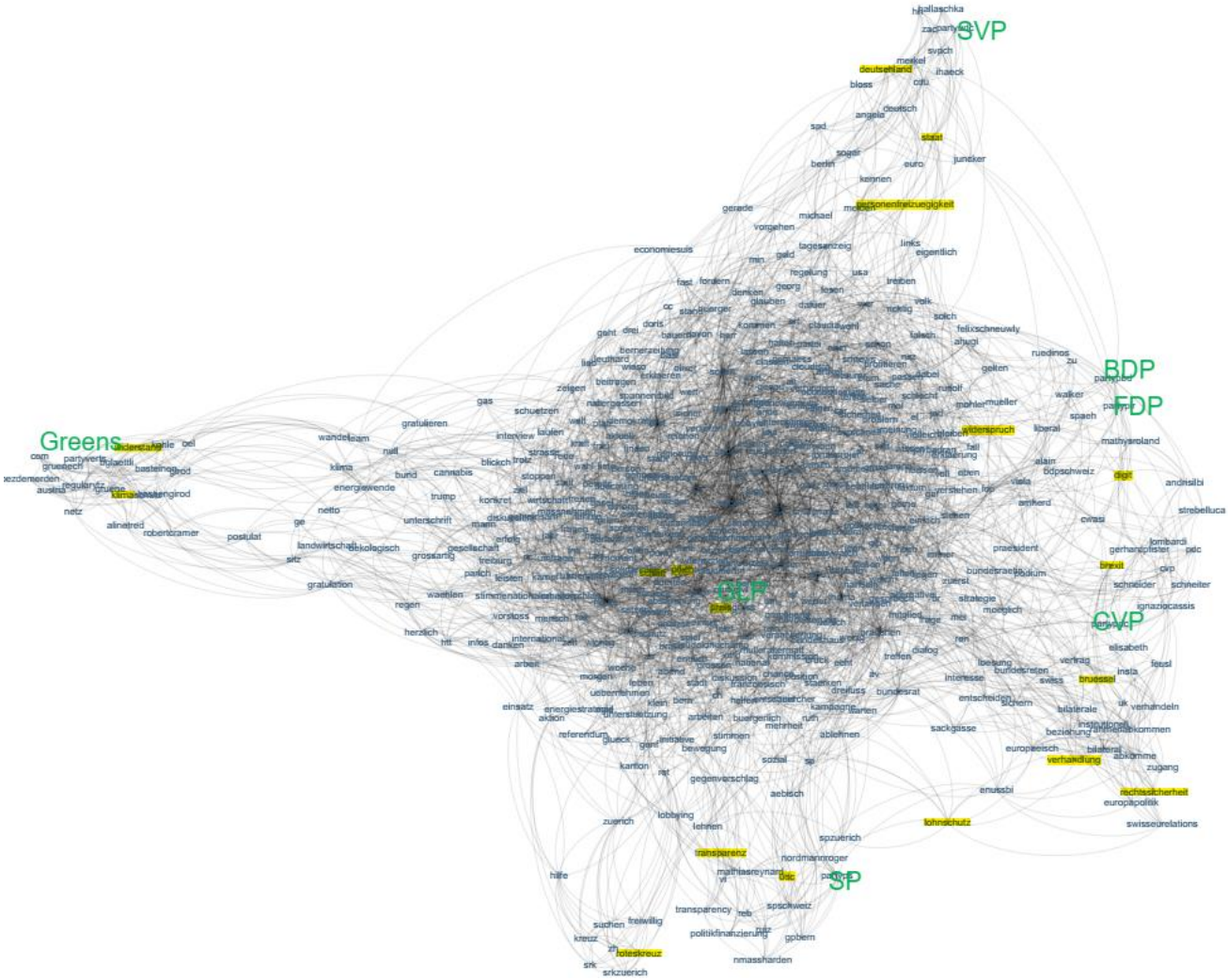


Figure 4: Semantic network from the EU-related tweets from parliamentarians.

Figure 4 demonstrates that partisan issue framing is more clear-cut on Twitter. For instance, the SVP emphasised concerns related to the freedom of movement (*personenfeizuegigkeit*) (especially, with respect to *Deutschland* 'Germany'), while the SP used Twitter to attack the views of the SVP and to focus the discussion on security (e.g., *transparenz* 'transparency' and humanitarian aid, e.g. *rotes kreuz* 'red cross') and social security (e.g., *lohnschutz* 'wage protection') related issues. As in parliamentary debates, the remaining parties were more concerned with regulation questions and how they could link the EU topic to their own agenda (e.g., climate, digitalisation, protection of rights). The global shape of the map reflects political opposites, in particular SP as opposite to SVP. While BDP and FDP appear very close, SP and Greens are not closely related in Figure 4. Their core issues do not seem to strongly overlap.

The results from the conceptual maps related to issue framing can also be assessed for their statistical significance. Table 3 displays the proportion of the top-words which entail the most statistically significant difference ($p < 0.001$) between the SVP and the other parties' communication about the topic "EU, Europe". We also used the two-proportions z-test for testing whether some words

are more prevalent (or "greater") in the SVP communication compared to the used of these words in the communication of the other parties. Table 3 mirrors the patterns from Figure 3 about the emphasis of the SVP on security and cost related issues in Parliament. Furthermore, Table 3 mirrors the Twitter trends (see Figure 4) where the SVP emphasise concerns related to the freedom of movement with a focus on Germany.

	SVP % (n)	Other parties % (n)
Twitter		
Merkel (Merkel)	0.75% (113)	0.14% (90)
Deutschland (Germany)	0.31% (47)	0.07% (42)
Brexit (Brexit)	0.3% (45)	0.12% (75)
Staat (State)	0.22% (34)	0.05% (33)
Personenfreizügigkeit (freedom of movement)	0.18% (27)	0.05% (29)
Euro (Euro)	0.16% (24)	0.05% (29)
Cdu (CDU)	0.15% (22)	0.03% (22)
Farage (Farage)	0.14% (21)	0.01% (6)
Juncker (Juncker)	0.14% (21)	0.03% (18)
Linke (Left)	0.14% (21)	0.04% (24)
Parliament		
Gpk (GPK)	0.3% (171)	0.22% (574)
Sicherheit (Security)	0.17% (96)	0.09% (227)
Kosten (Costs)	0.15% (84)	0.07% (184)
Rahmenabkommen (Institutional agreement)	0.11% (64)	0.07% (171)
Geschäftsprüfungsdelegation (Audit delegation)	0.1% (58)	0.06% (153)
Bundeskanzlei (Federal Chancellery)	0.09% (53)	0.04% (103)
Automatisch (Automatic)	0.08% (43)	0.04% (95)
Demokratie (Democracy)	0.07% (40)	0.03% (69)
Voting (Voting)	0.07% (38)	0.03% (69)
Gefahr (Danger)	0.06% (34)	0.03% (69)

Table 3: Proportion of the top-words which entail the most statistically significant difference ($p < 0.001$) between the SVP and the other parties' communication about the topic "EU, Europe".

5. Conclusions

Depending on the method used to assess issue ownership and issue framing, that is *what* politicians emphasise in discourse and *how* they speak about it, we arrive at a more nuanced picture when combining methods of distributional semantics. From the view of digital humanities, the proposed analytical strategy combining TM and KDE contributes to the development of computational approaches such as text mining and information visualisation, which are important to return patterns from a specific set of texts as accurately and effectively as possible (Tyrkkö 2020). The proposed approach provides detailed analyses of issue ownership and issue framing, while also offering transparency and focusing on new patterns which "can potentially overcome some of the shortcomings of both unaided humanistic interpretation and coding-based approaches" (Pääkkönen & Ylikoski 2021).

Whereas TM allows us to highlight the most salient topics and their prevalence according to

party affiliation, KDE is better suited to assess parties' framing of policy issues by looking at individual terms used to describe an issue. Using meta-information about the party affiliation and the main prevalent topic of each document further allows us to examine the semantic connection between parties and particular policy issues. Statistical tests were also useful to validate the content of the conceptual maps.

Our study entails several limitations worth addressing in future studies. First, temporal dimensions (such as years or decades) could also be included as a meta-information. Second, as we pointed out earlier, Twitter is only one possible social media source from which to extract data. Furthermore, we only looked at one country, Switzerland. Moreover, the size of the Swiss Twitter user sphere is comparatively moderate compared that of neighbouring countries (Müller, Wüest & Willi 2016). Cross-national comparisons are relevant as different uses of Twitter can prevail across countries. For instance, Swiss politicians relying on social media tend to use them in a unidirectional way, implying that they share partisan slogans or buzzes intended to attract media attention, but no real interaction with the broader audience to exchange ideas and engage in intensive debates about prominent issues occurs (Rauchfleisch & Metag 2016; Keller & Königslöw 2018). Third, we could have included additional pre-processing steps, such as the removal of mentions and actor names. Indeed, as the relationships among (political) users tend to be systematically clustered along ideological and/or socio-structural lines (Barberá & Rivero 2015; Barberá 2015), this could influence the shape of conceptual maps.

There are also future paths for research. For instance, we have used TM and KDE, but complementary methods could be envisaged, such as WordFish for the political scaling of politicians and parties on the left-right continuum (Grimmer & Stewart 2013). Furthermore, we have looked only at political discourse, other actors could be included in the analysis, such as media and citizens, who might also display selective attention to policy issues (Wüest 2018).

6. Acknowledgements

This paper was partly supported by the UFSP Digital Religion(s) project at the University of Zurich[24].

7. References

Arend, L. (1999). *Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries*. New Haven, CT: Yale University Press.

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political analysis* 23 (1), 76-91.

Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users. *Social Science Computer Review* 33 (6), 712–729.

Barberá, P., Casas, A., Nagler, J., Egan, P. J., Bonneau, R., Jost, J. T., & Tucker, J. A. (2019). Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review* 113 (4), 883–901.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics* 36 (4) 673–721.

Bélanger, E., & Meguid, B. M. (2008). Issue salience, issue ownership, and issue-based vote choice. *Electoral Studies* 27 (3), 477–491.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55 (4), 77–84.

Card, D., Boydstun, A., Gross, J. H., Resnik, P., & Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Short Papers, 2, 438–444.

[24] https://www.digitalreligions.uzh.ch/

Castanho Silva, B., & Proksch, S. (2021). Politicians unleashed? Political communication on Twitter and in parliament in Western Europe. *Political Science Research and Methods* 1–17. doi:10.1017/psrm.2021.36
Church, K. (2000). Empirical estimates of adaptation: The chance of two noriegas is closer to p/2 than p2. In *Proceedings of the 18th International Conference on Computational Linguistics*, COLING 2000, 1.
Curran, B., Higham, K., & Ortiz, E. (2018). Look Who's Talking: Bipartite Networks as Representations of a Topic Model of New Zealand Parliamentary Speeches. PloS One, 13 (6), e0199072.
Engesser, S., Ernst, N., Esser, F., & Büchel, F. (2017). Populism and social media: how politicians spread a fragmented ideology. *Information, Communication and Society* 20 (8), 1109–1126.
Enli, G. S., & Skogerbø, E. (2013). Personalized campaigns in party-centred politics: Twitter and Facebook as arenas for political communication. *Information, communication & society* 16 (5), 757–774.
Entman, R. M., Matthes, J., & Pellicano, L. (2009). Nature, sources, and effects of news framing. In K. Wahl-Jorgensen & T. Hanitzsch (Eds.), *The handbook of journalism studies*. London: Routledge, 195–210.
Ernst, N., Engesser, S., Büchel, F., Blassnig, S., & Esser, F. (2017). Extreme parties and populism: an analysis of Facebook and Twitter across six countries. *Information, Communication & Society* 20 (9), 1347–1364.
Evans, J. (2022). From Text Signals to Simulations: A Review and Complement to Text as Data by Grimmer, Roberts & Stewart (PUP 2022). *Sociological Methods & Research*, 00491241221123086.
Eve, M. P. (2022) *The Digital Humanities and Literary Studies*. Oxford: Oxford University Press.
Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. *Studies in Linguistic Analysis*, 1–32.
Gilardi, F., & Wüest, B. (2018). "Text-as-Data Methods for Comparative Policy Analysis." Working Paper. https://www.fabriziogilardi.org/resources/papers/Gilardi-Wueest-TextAsData-Policy-Analysis.pdf
Gilardi, F., Gessler, T., Kubli, M., & Müller, S. (2022). Social media and political agenda setting. *Political Communication* 39 (1), 39–60.
Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. New York, NY: Harper & Row.
Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis* 21 (3), 267–297.
Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.
Jacobs, K., & Spierings, N. (2019). A populist paradise? Examining populists' Twitter adoption and use. *Information, Communication & Society* 22 (12), 1681–1696.
Jacomy, M., Venturini, T., Heymann, S., Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software, *PLoS ONE* 9 (6), e98679.
Jones, T., Doane, W., & Jones, M. T. (2021). *"Package 'textmineR'. Functions for text mining and topic modeling". R CRAN package*. https://mran.revolutionanalytics.com/snapshot/2019-04-07/web/packages/textmineR/textmineR.pdf
Keller, T. R., & Kleinen-von Königslöw, K. (2018). Followers, spread the message! Predicting the success of Swiss politicians on Facebook and Twitter. *Social Media+ Society* 4 (1), 2056305118765733.
La Mela, M., Norén, F., & Hyvönen, E. (2022). Digital Parliamentary Data in Action (DiPaDA 2022): Introduction. In *Digital Parliamentary Data in Action (Dipada)* 3133, 1–8.
Mazzoleni, G., & Bracciale, R. (2018). Socially mediated populism: the communicative strategies of political leaders on Facebook. *Palgrave Communications* 4 (1), 1–10.
McClure, D. (2015). "Textplot refresh – Python 3". PyPI, CLI app.
Mollin, S. (2007). The Hansard hazard: gauging the accuracy of British parliamentary transcripts. *Corpora*, 2 (2), 187–210.
Moretti, F. (2013). *Distant Reading*. London: Verso.
Müller, C., Wüest, B., & Willi, T., (2016). Twitter data for political analysis in Switzerland. *Swiss Political Science Association Annual Meeting*, Basel, 21–22 January 2016. https://www.zora.uzh.ch/id/eprint/150398/1/ZORA150398.pdf
Pääkkönen, J. & Ylikoski, P. (2021). Humanistic Interpretation and machine learning. *Synthese* 199,1461–1497.
Petrocik, J. (1996). Issue Ownership in Presidential Elections, with a 1980 Case Study. *American Journal of Political Science* 40 (3), 825–50.
Rauchfleisch, A., & Metag, J. (2016). The special case of Switzerland: Swiss politicians on Twitter. *New Media & Society*,18 (10), 2413–2431.
Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of WSDM'15*, Shanghai, China.
Sahlgren, M. (2006). *The Word-Space Model: Using distributional Analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University.

Schneider, G. (2020). Changes in Society and Language: Charting Poverty. In P. Rautionaho, A. Nurmi & J. Klemola (eds.), *Corpora and the Changing Society: Studies in the Evolution of English*. Amsterdam: Benjamins, 29–56.
Schneider, G. (2022). Medical topics and style from 1500 to 2018. In T. Hiltunen a& I. Taavitsainen (eds.), *Corpus pragmatic studies on the history of medical discourse*. Amsterdam: Benjamins, 49–78.
Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics* 24 (1), 97–123.
Schwarzbözl, T., & Fatke, M. (2017). Measuring Populism in Social Media Data. *ECPR General Conference. Panel Measuring Populism and Populist Attitudes*, 6-9 September 2017, Oslo.
Sides, J. (2006). The Origin of Campaign Agendas. *British Journal of Political Science*,36 (3), 407–36.
Søyland, M. (2022). Party Control and Responsiveness. How MPs Use Variation in Lower-Level Institutional Design as an Electoral Responsiveness Mechanism. *Digital Parliamentary Data in Action Worksho.* Sweden: Uppsala. March 15, 2022. http://ceur-ws.org/Vol-3133/paper07.pdf
Stier, S. (2016). Partisan framing of political debates on Twitter. *Proceedings of the 8th ACM Conference on Web Science*. New York: Association for Computing Machinery, 365–366.
Taavitsainen, I. & Hiltunen, T. (2019). *Late Modern English Medical Texts: Writing medicine in the eighteenth century*. Amsterdam: Benjamins.
Tyrkkö, J. (2020). The war years: Distant reading British parliamentary debates. In J. Hansson & J. Svensson (eds.), *Doing Digital Humanities: Concepts, Approaches, Cases*. Växjö: Linnaeus University Press, 169–199.
Van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., & Beunders, H. (2017). The debates of the European Parliament as linked open data. *Semantic Web* 8 (2), 271–281.
Wüest, B. (2018). Selective Attention and the Information Environment: Citizens' Perceptions of Political Problems in the 2015 Swiss Federal Election Campaign. *Swiss Political Science Review* 24 (4), 464–486.
Zucchini, W. (2003) *Applied smoothing techniques. Part 1: Kernel density estimation*. Unpublished manuscript. http://staff.ustc.edu.cn/~zwp/teach/Math-Stat/kernel.pdf

Learning Inductive and Deductive Topics in Parallel Using Seeded Topic Modeling

Patrick Kahle, Fritz Kliche

Universität Bielefeld, Universität Hildesheim

E-mail: patrick.kahle@uni-bielefeld.de

Abstract

Linking empirical material with existing theory confronts (social) scientists either with problems of inductive approaches or with such of deductive approaches. We use different topic modeling methods as a bridge between the two poles. We work on a corpus of interviews with representatives of organizations involved in migration and focus on their perceptions of solidarity towards refugees.

On the one hand, we examine arguments which can be attributed to what Boltanski and Thévenot (2006) would call "Justification Worlds" (JWs) – e.g., civic or humanitarian motives for the interviewee's engagements. The foundational texts of JWs suggest a fixation on central terms which are given *ex ante*. This would represent dictionary-based coding. Deductive hypothesis-testing research designs are affected by the objections to deductive procedures (cf. Gibbon 2010), as merely tying the previously given concepts together, applying them on a corpus, and tightening the lasso exerts violence on the material and obscures field- and material-specific features. On the other hand, inductive approaches such as topic modeling reveal the content topics of a corpus, but the results have to be bound to existing theory.

Aiming at linking both concepts, we use topic modeling for finding words which indicate the contents of our corpus and odd out words which can be attributed to the JWs 'industrial' and 'civic', as well as words which are indicative for 'higher-level', generic topics such as 'refugees' or 'organizations'. We use these words as seeds for 'seeded topic modeling' and train a model along with additional unseeded topics. We find that the latter indicate topics not found in an unseeded topic model, and even reveal an additional JW: 'domestic'.

Keywords: seeded topic modeling, justification worlds, economics of convention, distant reading, evaluation of topic modeling

1. Introduction

Social sciences distinguish between inductive and deductive approaches for the analysis of empirical data. Both have to be applied carefully when used for providing evidence for a specific theory. This is the case here, too. We work on interview data collected in Germany in the aftermath of the so-called refugee crisis 2015–2016. The interviewees are organized people acting in solidarity with refugees. Besides this common ground, we find different motivations and conventions for the social commitment. The arguments which are raised are not arbitrary and based on existing discourses – Sellars (2007) or Brandom (2000) would probably call them "spaces of reasons". The Economics of Convention (EC) and especially the justification logics (or "Justification Worlds", JWs) proposed by Boltanski and Thévenot (2006) offer an approach to quantify such conventions. They argue that these worlds are "sufficient to describe justifications performed in the majority of ordinary situations", adding that they are "historical constructions" which can change over time (Boltanski & Thévenot 1999, 369). They form the semantic network of the worlds of justification, a linguistic condensation, so to speak, of the worlds conceived as materialistic in themselves (Boltanski & Thévenot 2007, 222–286). The foundational texts of EC suggest a fixation on central terms, therefore a dictionary-based approach to refer to these worlds seems plausible.

Contrary to the suggestion of the basic EC texts, we argue that JWs should not be determined conclusively and therefore deductively in case of our interviews. EC is less fixated on terms than on concepts; accordingly, we integrate an inductive approach. We begin with the vocabulary of the JW dictionaries which occurs in the interview corpus and validly represents the concepts of the JWs from our point of view. Then we adapt this vocabulary to the content of our corpus using keyATM (Eshima, Imai & Sasaki 2020) as an implementation for seeded topic modeling. Our contribution is a case study

which focuses on methodology. We see the detection of theoretical cultural objects in text data as postulated by Thévenot and Boltanski by means of algorithms as our primary goal. We elaborate two JWs which became apparent in our text material after topic modeling – 'industrial' and 'civic' – not all JWs proposed by Thévenot and Boltanski.

Topic modeling based on Latent Dirichlet Allocation (LDA) is well established for quantitative discourse analyses. E.g., Ylä-Anttila & Luhtakallio (2016) and Ylä-Anttila, Eranti & Kukkonen (2022) use LDA for frame analyses to propose an analytical framework, Justifications Analysis (JA). But in their JA, the codes are not identified inductively from the data but based on justification principles. Because these principles are relatively stable, the approach, according to these authors, should be particularly suitable for comparative research. Ylä-Anttila and Luhtakallio (2016) expect that "factors such as media logics, editorial decisions and sourcing practices have an influence on the types of frames that are linked to different speaker groups." We follow this approach with comparable applications to interviews.

Topic models are trained on words which tend to co-occur in documents. Such word-based approaches have difficulties with low frequency words. As our corpus is relatively small and covers a broad range of narratives, it is prone to this problem. There are proposals to address this issue. Dieng, Ruiz & Blei (2020) do not apply LDA on words but on word embeddings, i.e. vector representations of words, which has the advantage that the topic attributions of words with low frequencies are also determined by similar, but more frequent words. Bianchi, Terragni & Hovy (2021) use contextualized word embeddings where the vector representation of a word also depends on its context, in their case, the sentence in which a word occurs. We still use a word-based implementation as it offers the possibility of seeded topic modeling.

In section 2, we introduce our corpus data and other resources. In section 3, we summarize the objections on dictionary-based approaches especially according to JWs. Our approach is described in section 4. We present and discuss the results in sections 5 and 6 and conclude in section 7.

2. Data and Resources

We use a corpus of 53 interviews (about 423.000 words) which were conducted in German in 2019/2020. The interviews were transcribed by student employees and checked by one of the authors. Since digital methods are to be used for the analysis of the interviews, mandatory errors and dialects were adjusted to the written language. We only process the parts of the interviewees. They are representatives of governmental organizations (offices for migration and integration) and social welfare organizations, working with migrants and refugees during and after the so-called refugee crisis (2015 ff.). A contribution of an interviewee in our corpus may look like the example in (1)

(1) *Richtig. Ich mache Quartiersarbeit. Ziemlich viel auch ja. Oder unser Bücherschrank, der wird hier auch hauptsächlich von den Anwohnern genutzt und nicht so stark von den Geflüchteten und so [Laut] ne.*
'Right. I do community work. Pretty much yes. Or our bookcase which is mainly used here by the residents and not so strongly by the refugees and so [filler] ne.'

Our second resource is an existing EC dictionary (Middelschulte & Kahle 2019). It is based on the terms given by Boltanski & Thévenot (2007) as indicators of the (six original) worlds of inspiration, renown, civic, market, industrial and domestic, and was supplemented by terms of the later added project-oriented world and the ecological world. All terms that Boltanski and Thévenot collected and highlighted as representative of the respective worlds were included in the dictionary which in sum

comprises of 1.373 terms. Some entries which are commonly stop words from a natural language processing (NLP) perspective were removed by Middelschulte & Kahle (2019), who successfully applied the dictionary on curricula of school subjects related to social sciences.

Third, Eshima, Imai & Sasaki (2020) present the R package keyATM (keyword-assisted topic modeling). Similar to unsupervised topic modeling using LDA (cf. e.g., Blei, Ng & Jordan 2003), the number of 'topics' to be distinguished has to be predefined by the user. Fitting to the thought of liminal objects, topic modeling allows the occurrence of terms in multiple topics. One of the outputs of LDA are the *n* most significant words for each topic. In seeded topic modeling, words can be attributed to topics before training. These 'seed words' bias the generation of the topic clusters during training. keyATM allows to bias some topics and to include additional unseeded topics.

3. Dictionary-Based Approaches on JWs

JWs are worlds which are common to all of us because they are public (*in sensu* Boltanski & Thévenot 2006; also cf. Ylä-Anttila & Luhtakallio 2016). Nevertheless, such 'worlds' do not simply derive from social ideas but are results from discourses long handed down in social and cultural history. They arise from permanent social contradictions and can be conceived as pacification strategies of permanently smoldering crises. One of the most important functions of the common worlds is to make the observation of social inequality tolerable. The non-arbitrariness of reasons now raises the question of whether these worlds can be drawn upon arbitrarily in any social situation, or whether there are structural features of the social situation that favor connection to a specific JW in each case (Honneth 2010). This, too, calls for quantification and comparison with other variables – such as different words constituting a JW, or JWs used in different social spheres.

Boltanski and Thévenot (2006) and other classical texts on EC assume that JWs are manifested in central terms, which motivates a dictionary-based approach. The criticism in the discussion of methods according to which coding instructions and code books for human coders are often already available, in contrast to dictionaries (Landmann & Züll 2004, 119; Scharkow 2011, 549–550), does not apply in our case: there exists an EC dictionary (Middelschulte & Kahle 2019). Initially, we could assume that an analysis of a corpus in which we apply a given theory (EC) with its own dictionary would be sufficient. This would represent a deductive hypothesis-testing research design – more precisely, dictionary-based coding (Scharkow 2011, 548). A text mining software generates a term-document matrix recording the occurrences of terms in a unit of analysis. In this dictionary-based approach, the relevant terms are defined *ex ante* (Hippner & Rentzmann 2006, 289; Scharkow 2011, 548).

But dictionary approaches face – besides other discussions on preprocessing steps – the well-known critique that complex coding rules would still be undermined by classical deterministic dictionary coding which proceeds in a strictly word-oriented way – a category *c* is assigned if word *w* is contained in the coding unit – and that it would thus require a relatively large effort to develop a sophisticated and valid dictionary with complex coding rules. For this purpose, weighted word lists and negative keywords could be developed in the sense of probabilistic coding (Scharkow 2011, 549). A theory such as EC, which moreover exhibits even different axioms per JW, makes close reading necessary. Blended reading is the standard here (Dumm & Niekler 2016, 91). Additionally, EC knows besides specific (world-related) entities liminal entities (Dodier 1993) undermining the uniqueness provided by dictionaries which is a prerequisite of a dictionary approach.

In the following, we argue methodologically why we combine our dictionary-based and thus deductive approach with an inductive component which is based on seeded topic modeling. Each text

can be understood as an assembly, a multidimensional gathering of different levels or axes of sense (*in sensu* Minsky 1975; Ziem 2008). Texts contain factual statements, psychological suggestions, moral justifications, and so forth. We assume that JWs cut across content topics. We see our corpus as multidimensional space in which topics, JWs and regimes of engagement (Eranti 2018; Luhtakallio & Thévenot 2018) are arranged on different axes alongside others, but are intertwined in narratives and therefore cannot be recorded distinctly, e.g., by using unsupervised topic modeling.

4. Method: Application of an EC Dictionary to Seeded Topic Modeling

Our approach is to find these dimensions in parallel. First, we use keyATM for unsupervised topic modeling which retrieves the most significant words for topics covering the content of the empirical field. We find that these topics can be grouped into four superordinate categories: (1) problems and issues of refugees, (2) the everyday life of the interviewees, (3) cooperations with local organizations and institutions, (4) and reflection on social and political issues in general. From this result, we extract strongly indicative keywords for these superordinate topics, as well as keywords which can be attributed to two relevant JWs in our field of research: 'civic' and 'industrial'. Next, we use keyATM for seeded topic modeling. Using the keywords extracted from the first result, we train a model with six biased topics (the four superordinate topics and both JWs). This procedure is suitable for preserving the superficial theme dimension but bending it so far that other dimensions emerge as axes (topics). The method is thus reminiscent of Principal Component Analysis (Pearson 1901).

The implementation of this approach requires several NLP preprocessing steps. First, we exclude all texts from the interviewers (the 'questions') and segment the corpus into 2831 paragraphs as the units to be assigned with topic probabilities. The texts are lemmatized using the TreeTagger (Schmid 1995) and an additional postprocessing application of Gojun et al. (2012) which heuristically assigns lemmata to tokens which are unknown to the tagger. We delete all words which are not labeled as nouns, adjectives, adverbs or verbs. The result of our first run of keyATM without seeded topics were the 20 most significant words for 20 topics. As they contained words which hardly contribute to – as thus dilute – the detection of coherent topics (e.g., *be*, *have*, *say*), we deleted 48 of these words from the corpus as stop words, which are listed in the appendix 1. The filterings reduced the corpus from originally ~570.000 tokens to ~87.000 words. Last, all remaining words were converted to lowercase. (2) and (3) show how these preprocessing steps 'sharpen' the input data, as the albeit much fewer remaining words are semantically more relevant for the detection of topics.

(2) *Also Sie befinden sich hier im Organisation276. [Laut] Ähm das Projekt heißt Projekt72. Wird gefördert von der Organisation68 beziehungsweise von dem Ort18 Ministerium für Schisslaweng. (lachen) [Laut] Ähm wir sind hier ein Gemeinwesenprojekt zur Hälfte.*
‘So, you are here at Organization276. [filler] Ehm, the project is called Project72. Is funded by Organization68, respectively, by City18 ministry-of-something. (laugh) [filler] Ehm, we are here a community project, halfway.’

(3) *befinden projekt projekt gefördert ministerium Schisslaweng gemeinwesenprojekt hälfte*
‘be-located project project fund ministry some-place community-project halfway’

4.1. First Run: Unseeded Topic Modeling

Maier et al. (2018) discuss methods for the application and evaluation of (LDA-based) topic modeling. They point to the vagueness of the concept "topic" and that it remains unclear what a "topic" actually

means (ibid., 3). Usually, topic models with different numbers of topics are trained. With a higher number of topics to be detected, the topics become more specific (ibid., 11). We first trained different topic models without using seed words. After testing for 6 up to 20 topics and evaluating the 20 most significant words for each topic, it turned out that 15 topics led to a model for which the topics could be given the most meaningful and coherent interpretation.[25] We found that these topics could be subsumed under four superordinate categories:

Refugees:	*Topics which address the living conditions and needs of refugees*
We:	*The interviewees portray their own work and working conditions.*
Organization:	*Topics which concern the cooperation with the local administration*
Politics:	*Matters which are not descriptions of the daily lives of the interviewees or refugees, but descriptions of the current situation, e.g., referring to solidarity of the society in general*

Table 1 shows excerpts of the results – the complete table with the 20 most significant words for the 15 topics, as produced by the algorithm, and the categorization of the topics into the four superordinate topics, which is based on our interpretation, is given in the Appendix 2. The colors indicate the categories 'refugees' (green); 'we' (blue); 'organization' (yellow); and 'politics' (orange). We collected from these results words which we see as indicative for the superordinate categories for subsequent use as seed words – again, based on subjective interpretations of the results. They are printed in bold script in the table.

Topic 1	mensch 'human'/'person', stehen 'to stand', mann 'man', bekommen 'to get', halten 'to hold'/'to stop', leben 'live'/'to life', **aktion 'action'**, stadt 'town'/'city', **schreiben 'to write/writing'**, gespräch 'conversation', bringen 'to bring', lesen 'to read', gehören 'to belong', **politik 'politics'**, […]
Topic 2	**leute 'people'**, arbeiten 'to work', gut 'good', landkreis 'county', arbeit 'work', unterkunft 'accommodation', total 'total', sprechen 'to speak', **persönlich 'personal'**, privat 'private', klein 'small', wahrscheinlich 'probable', einzeln 'single'/'separate', schön 'beautiful', politisch 'political', […]
Topic 3	**projekt 'project'**, ehrenamtliche 'voluntary', leute 'people', **sache 'matter/thing'**, jahr 'year', **landkreis 'county'**, gut 'good', **stelle 'office/job'**, geflüchtete 'refugees', arbeiten 'to work', **stadt 'town/city'**, treffen 'meeting/to meet', teil 'part', sitzen 'to sit', zeit 'time', […]
Topic 4	jahr 'year', leute 'people', anfang 'beginning', bündnis 'alliance', mann 'man', klein 'small', arbeit 'work', tisch 'table', kümmern 'to take care', aktiv 'active', **stadt 'town/city'**, **sache 'matter/thing'**, rund 'round', laufen 'walk'/'be in process', veranstaltung 'event', aktion 'action', […]
Topic 5	leute 'people', **geflüchtete 'refugees'**, arbeit 'work', problem 'problem', freiwillige volunteers', **brauchen 'to need'**, cool 'cool', **sachen 'matters/things'**, helfen 'to help', **deutsch 'German'**, anfang 'beginning', relativ 'relative', voll 'full', begeistern 'to inspire'/'thrill', erfahrung 'experience', […]
[…]	[…]

Table 1: Results from unseeded topic modeling, shortened. The complete table is given in the Appendix 2.

Some words appear in several topics, including different superordinate topics, which is a prominent feature of topic modeling and comes close to the idea of liminal entities (Dodier 1993). By selecting such words as seeds, we bias them towards one specific topic, e.g., "refugees" or "association".

Next, we defined seed words for the JWs 'industrial' and 'civic'. They are based on the

[25] We found an interpretation for every topic, which is unusual for topic models but more likely in view of the multiple preprocessing steps. These and all other interpretations were carried out by both authors, the first knew the material as an interviewer, the second joined as a computational linguist.

interpretation of the results of unseeded topic modeling, as it reveals words which are significant for our corpus and thus should not be omitted, but which are also in the lists of Middelschulte and Kahle (2019). We only selected generic words as seeds which would not bias the seeded topics towards the vocabulary of a specific narration or interviewee, examples are "structure", "society", "to speak", "contact". Table 2 summarizes the seed words for the four superordinate topics and the two JWs.

Politics	situation 'situation', politik 'politics', kultur 'culture', aktion 'action', schreiben 'to write/writing', abschiebung 'deportation', abschieben 'to deport'
Refugees	sprechen 'to speak', deutsch 'German', bekommen 'to receive/get', lernen 'to learn', arbeiten 'to work', brauchen 'to need', sachen 'matters/things', geflüchtete 'refugees'
Organization	projekt 'project', landkreis 'county', stadt 'town/city', kommune 'municipality', kollege 'colleague', sache 'thing', struktur 'structure', kollegin 'female colleague', stellen 'offices'/'jobs'/'to provide', stelle 'office'/'job'
We	kontakt 'contact', verein 'association/club', treffen 'to meet', leute 'people', engagieren 'to be committed'/'to be engaged'/'to engage', persönlich 'personal', kennen 'to know', stadtteil 'quarter/neighborhood'
Industrial	zeit 'time', schaffen 'to accomplish', funktionieren 'to be working/functioning', organisieren 'to organize', struktur 'structure', rund 'round', laufen 'to walk/be in process', bereich 'domain/area'
Civic	solidarität 'solidarity', verein 'association/club', gleich 'equal'/'soon', politisch 'political', solidarisch 'solidary', gesellschaft 'society', dürfen 'to be allowed'

Table 2: Seed words for four superordinate topics and two JWs.

4.2. Second Run: Seeded Topic Modeling

Using the seed words for six seeded topics, we trained several models which differ in the number of additional unseeded topics. We ran experiments for nine models, contributing 4, 5, 6, 7, 8, 9, 10, 20 and 40 additional topics.[26] The goal of our analysis is to maximize the coherence within the *seeded* topics; for similarly good models, the coherence of the unseeded topics is then considered in a second step of the evaluation, using human judgement approaches resp. eye balling models. We evaluate in two ways: we judge the seeded topics observation-based – in concrete, we observe among the 20 most significant words those which indeed indicate the seeded topics. The unseeded topics we judge interpretation-based – in concrete, we observe words intruding a coherent topic, in alignment with the "intrusion test" proposed by Chang et al. (2009, 3 f.). Due to space constraints, only the first part of the evaluation is comprehensively described and illustrated in this paper – in other words, the full consideration between the two best models is not discussed here.

For the first evaluation (the seeded topics), we make use of keyATM's highlightings of those words which were previously given as seeds for finding the parameter for which the seeded topics have the strongest bias by the seed words. We use (1st) the total number of seed words among the 20 most significant words in each seeded topic; (2nd) the sum of the inverted ranks of the seed words, whereby seeds with a higher rank get a higher value; and (3rd) the average ranked sum, which ensures that a high rank total is not just the result of a high number of seeds in the top 20, but reflects the position of the seed words in the top 20.

[26] In the following, we name the different models corresponding to the number of additional topics. E.g., model M-7 consists of the six seeded topics and seven additional unseeded topics.

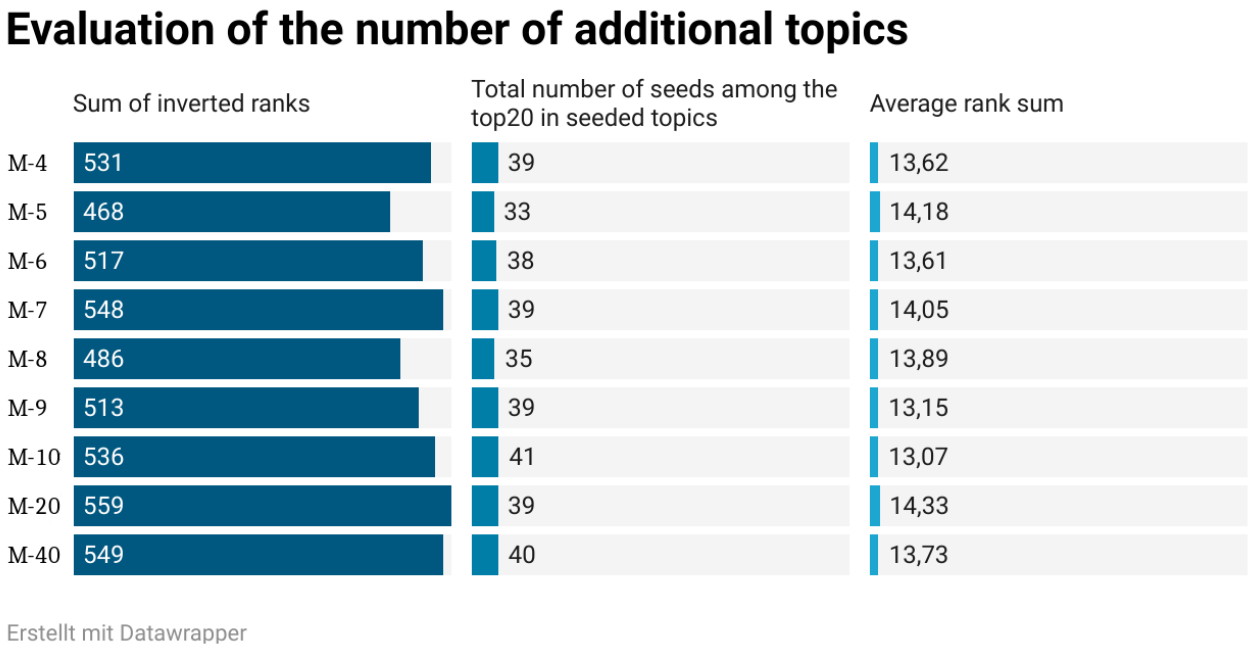


Figure 1: Evaluation of the number of additional topics.

If we take these indicators into account, we arrive at the following evaluation (Figure 1). Based on the first indicator, there are only marginal differences, but models 5 and 8 perform comparatively poorly here. This is also reflected in the sum of the inverted rank. Models 4, 7, 10, 20 and 40 are particularly strong. If we consider the average rank sum, models 4 and 40 are (relatively) poor and model 10 is very poor. Therefore, the models in which we use 7 and 20 additional topics (M-7 and M-20) score best in our evaluation. They are the groundwork of our subsequent validation of the outputs, evaluating if the word clusters actually represent what we think they do (DiMaggio, Nag & Blei 2013; Grimmer and Stewart 2013; Evans 2014).

5. Results

First, we analyzed the results of M-7. Table 3 shows the most significant words for one seeded topic – the JW 'industrial' – and excerpts of the most significant words of the 7 additional topics, again using background colors which are based on our interpretation.

In green, we highlighted unseeded topics for which we could recognize a clear meaningful interpretation, which are:

Topic 1: *(Problems of) housing and hosting*
Topic 3: *(Rights of) groups of refugees*
Topic 4: *(Emotional) interactions of refugees and their supporters*
Topic 5: *The domestic world*
Topic 7: *Organization of work and daily work*

Interestingly, in topic 5 we identify the 'domestic world', one of the JWs in Boltanski and Thévenot (2006). This topic is not found among the unseeded topics given in table 1.

In the sense of traffic lights, we highlight in yellow topics 2 and 6 for which we could find an interpretation, but not a generic and meaningful topic for the contents of our corpus and the aims of our study. Several words for topic 2 can be attributed to the concept of 'waiting', be it at a doctor, or are words which refer to the concept 'time' like "clock", "week", "days", "hours", "quarantine", as

contained in the stories of the interviews. Topic 6 is a compilation of terms found for a large part in one interview, as "exhibition", "ship", "photograph" and "captain" refer to stories of this particular interview.

Industrial	bereich 'domain'/'area' ✓, good 'gut', zeit 'time' ✓, organisieren 'to organize' ✓, gemeinsam 'together', schaffen 'to accomplish' ✓, jahr 'year', rund 'round' ✓, funktionieren 'to be working'/'to be functioning' ✓, unterstützen 'to support', gemeinde 'community', kirche 'church', unterschiedlich 'different', tisch 'table', bestimmt 'certain', ehrenamtliche 'volunteers', klein 'small', problem 'problem', tischen 'tables', laufen 'to walk'/'to be in process' ✓
unseeded Topic 1	wohnung 'apartment', unterbringen 'to accommodate', privat 'private', wohnraum 'living space'/'housing space', problem 'problem', familie 'family', sozial 'social'/'societal', wohnen 'to live'/'to lodge', deutschland 'Germany', n 'n', dezentral 'decentralized', verhindern 'to prevent', kennen 'to know' [4], block '(apartment) block', […]
unseeded Topic 2	mensch 'human'/'person', uhr 'clock', woche 'week', tagen 'days', englisch 'English', stunden 'hours', fahren 'to drive', politisch 'political' [6], karte 'map', arzt 'doctor', sitzen 'to sit', heimat 'home (country)', versorgung 'supply', behandeln 'to treat', tee 'tea', quarantäne 'quarantine', […]
unseeded Topic 3	arbeiten 'to work' [2], sache 'matter'/'thing' [3], jugendliche 'youths', land 'land', beratungsstelle 'information center', frau 'woman', extrem 'extreme', grund 'reason', anspruch 'claim', kennen 'to know' [4], arbeitslosengeld 'unemployment benefit', […]
unseeded Topic 4	arbeit 'work', geflüchtet 'fled', jahr 'year', fragen 'to ask', gruppe 'group', freiwillige 'volunteers', kind 'child', jugendlich 'adolescent', emotional 'emotional', ort 'place', unglaublich 'unbelievable', gelten 'to be valid'/'be in force', erwartung 'expectation', […]
unseeded Topic 5	kind 'child', familie 'family', sache 'matter'/'thing' [3], mann 'man', reden 'to speak', toll 'great', eltern 'parents', verstehen 'to understand', ja 'yes', tochter 'daughter', interkulturell 'intercultural', lesen 'to read', frühstücken 'to eat breakfast', erzählen 'to tell', […]
unseeded Topic 6	bild 'picture', frau 'woman', ausstellung 'exhibition', passieren 'to happen', mann 'man', schicken 'to send', platz 'space'/'place', zweit 'second'/'in pairs', Schiff 'ship', kopftuch 'headscarf', foto 'photo', fotografieren 'to take a picture', erleben 'to experience', tragen 'to carry', freundlich 'friendly', werfen 'to throw', kapitän 'captain', […]
unseeded Topic 7	kollegin 'female colleague' [3], tag 'day', büro 'office', kurz 'short', zuständig 'responsible', e-mail 'email', türe 'door', telefon 'telephone', termin 'appointment', frage 'question', zweit 'second'/'in pairs', anrufen 'to call', handlungsfeld 'field of action', warten 'to wait', bitten 'to request', wohnung 'apartment', zeit 'time' [5], schicken 'to send', klient 'client', team 'team', […]

Table 3: Results from seeded topic modeling, shortened. The complete table is given in the Appendix 3. ✓ marks seed words; [number] refers to words of the unseeded topics which are also seed words for the seeded topics. The colors green and yellow indicate coherent and less coherent topics.

We also interpreted the results of M-20. We noticed that for these results, the 'significant words' are to be considered carefully, as they often correspond to a frequency in the corpus in a single-digit range. In some cases, the assumption of 20 topics tears apart topics. For instance, two of these topics share words such as "exhibition", "photograph" and "captain", which – we know – refer to the stories of a particular interview. On the other hand, topics can include several concepts ('catch-multiple') which appear only coherent when the content of the interview is known. Still, for 5 of the 20 additional/unseeded topics, we found a coherent and meaningful interpretation. For nine topics, we found a weaker interpretation. For six topics, it was unclear how to interpret them. We argue that M-7 yields generic topics, whereas the topics of M-20 are fine-grained and can often be attributed to one interview.

6. Discussion

The narrations of our interview partners constitute the language and vocabulary of the corpus. They refer to specific situations, such as the report of a photo exhibition in favor of refugees, which a captain visits, which explains the word cluster "exhibition", "ship", "photo", "to photograph" and "captain" in unseeded topic 6. Such specific contexts are in contrast to two kinds of generic 'topics' prevalent in our corpus: First, JWs are interwoven into the narrations and become apparent by generic words which are not attributed to a specific context. Second, we find from the unseeded topic models that topics can be subsumed under the general categories 'Politics', 'Refugees', 'Organization' and 'We'. Topics of both categories are found in a variety of texts and dominate the formation of topics in an unseeded approach. Still, these general topics are included in what we aim at, so, narrowing down the detection of topics to more specific topics would probably prevent the distant view on the whole corpus, but would instead cluster words as topics which constitute the vocabulary of a single narration. Also fine-tuning the hyperparameters of the topic model (George & Doss 2018) would hardly reveal these intertwined concepts.

So, our approach is to select generic words for the two JWs and the superordinate topics and use them as seeds for training a seeded topic model, which also includes additional *un*-seeded topics. Thévenot and especially Dodier (1993) also distinguish different levels of generality and specificity: entities or concepts not linked to any JW, entities linked to a specific JW, and entities which constitute links between JWs, e.g., "trade unions" which can be seen as link between the 'civic' and the 'industrial' world. Topic modeling allows the affiliation of terms to different topics, even if they are seeds of a topic. This accommodates the logic of general and specific entities of JWs.

We know that our setup has limitations and loose ends for further work. The biggest objection that could be raised is against the selection of seed words for the inductive *super*-ordinate topics. Our selection was based on knowledge of the interview material, the majority of which was collected by one of the authors and has been analyzed qualitatively. A critical objection could be raised that this is precisely what produces a bias, since analysts would have made more data-driven choices in ignorance of the specific interviews. Through our dialogic approach between an interviewer and a computational linguist, we aimed at least to reduce this bias.

Our experiments are based on the assumption that 'topics' are characterized by single words. While this is in alignment with the standard use of topic modeling (e.g., keyATM expects single tokens as 'seeds'), this approach is agnostic to indicators which go beyond single words, which could be addressed by e.g., vector representations of the words, possibly including context information, as mentioned in the examples in section 1.

A further step for our project would be an iteration in which we seed the domestic world. Next steps in sense of the method would be the application of quantitative metrics like held-out likelihood and coherence calculations additional to the human judgement (*in sensu* Lau, Newman & Baldwin 2014). In this way, a large number of iterations would conceivably optimize the results, even though M-7 already convinces us well.

7. Conclusion

In our experiments on seeded topic modeling, we combine inductive and deductive approaches for mapping concepts of JWs onto a specific context – a corpus of transcribed interviews with people engaged in support for refugees. We argue that we thus do not superimpose but adapt concepts which were given *ex ante* onto the topics which are immanent in our language domain.

We use topic modeling which reveals topically significant words in our corpus. From this result, we extract words indicating the JWs 'industrial' and 'civic', as well as words which are indicative for the generic topics 'politics', 'refugees', 'organization' and 'we'. We use them as seed words for training topic models with different numbers of additional unseeded topics. We find that models with seven (M-7) and twenty (M-20) additional topics are convincing with regard to our evaluation. M-7 yields generic topics and even another JW, whereas the topics of M-20 are (too) fine-grained and can often be attributed to one specific interview.

8. Acknowledgements

This paper is based on material of the research project 'Discourses on Solidarity in Times of Crisis: Analyzing and Explaining Solidarity in the Context of Migration' (online: https://www.uni-hildesheim.de/soldisk/en/soldisk), which was gratefully financially supported by the Ministry of Science and Culture of Lower Saxony. We thank our supervisors Prof. Dr. Michael Corsten and Prof. Dr. Ulrich Heid at the University of Hildesheim. We are also indebted to our two unknown reviewers and Jenny Tarvainen for editorial advice.

References

Bianchi, F., Terragni, S. & Hovy, D. (2021). Pre-training is a hot topic: contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLO 2021, Volume 2: Short Papers)*, online. Bangkok: Association for Computational Linguistics (ACL), 759–766.

Blei, D., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (1), 993–1022.

Boltanski, L. & Thévenot, L. (1999). The sociology of critical capacity. *European Journal of Social Theory* 2 (3), 359–377.

Boltanski, L. & Thévenot, L. (2006). *On justification. Economies of worth*. Princeton: Princeton University Press.

Boltanski, L. & Thévenot, L. (2007). *Über die Rechtfertigung. Eine Soziologie der kritischen Urteilskraft*. Hamburg: Hamburger Edition.

Brandom, R. (2000). *Articulating reasons. An introduction to inferentialism*. Cambridge: Harvard University Press.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S. & Blei, D. (2009). Reading tea leaves: how humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams & A. Culotta (eds.), *Advances in Neural Information Processing Systems.* Vancouver, 6–9 December 2009, 288–296.

Dieng, A., Ruiz, F. & Blei, D. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics* 8, 439–453.

DiMaggio, P., Nag, M. & Blei, D. (2013). Exploiting affinities between topic modelling and the sociological perspective on culture: application to newspaper coverage of U.S. government arts funding. *Poetics* 41 (6), 570–606.

Dodier, N. (1993). Les appuis conventionnels de l'action. Eléments de pragmatique sociologique. *Réseaux* 11 (62), 63–85.

Dumm, S. & Niekler, A. (2016). Methoden, Qualitätssicherung und Forschungsdesign. Diskurs- und Inhaltsanalyse zwischen Sozialwissenschaften und automatischer Sprachverarbeitung. In M. Lemke & G. Wiedemann (eds.), *Text Mining in den Sozialwissenschaften*. Wiesbaden: Springer, 89–116.

Eranti, V. (2018). Engagements, grammars, and the public: from the liberal grammar to individual interests. *European Journal of Cultural and Political Sociology* 5 (1–2), 42–65.

Eshima, S., Imai, K. & Sasaki, T. (2020). Keyword assisted topic models. Working Paper. arXiv:2004.05964

Evans, M. (2014). A computational approach to qualitative analysis in large textual datasets. *PLoS One* 9 (2), 1–10.

George, C. P. & Doss, H. (2018). Principled selection of hyperparameters in the Latent Dirichlet Allocation model. *Journal of Machine Learning Research* 18 (162), 1–38.

Gibbon, D. (2010). Lexika für multimodale Systeme. In K.-U. Carstensen, C. Ebert, C. Ebert, S. J. Jekat, R. Klabunde & H. Langer (eds.), *Computerlinguistik und Sprachtechnologie. Eine Einführung.* Heidelberg: Spektrum, 515–523.

Gojun, A., Heid, U., Weissbach, B., Loth, C. & Mingers, I. (2012). Adapting and evaluating a generic term extraction tool. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation.* May 23rd–May 25th 2012. Istanbul: European Language Resources Association (ELRA), 651–656.

Grimmer, J. & Stewart, B. (2013). Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21 (3), 267–297.

Hippner, H. & Rentzmann, R. (2006). Text Mining. *Informatik Spektrum* 29 (4), 287–290.

Honneth, A. (2010). *Das Ich im Wir. Studien zur Anerkennungstheorie*. Berlin: Suhrkamp.

Landmann, J. & Züll, C. (2004). Computerunterstützte Inhaltsanalyse ohne Diktionär? Ein Praxistest. *ZUMA Nachrichten* 28 (54), 117–140.

Lau, J. H., Newman, D. & Baldwin, T. (2014). Machine reading tea leaves: automatically evaluating topic coherence and topic model quality. In S. Wintner, S. Goldwater & S. Riezler (eds.), *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics.* April 26th–April 30th 2014. Gothenburg: Association for Computational Linguistics (ACL), 530–539.

Luhtakallio, E. & Thévenot, L. (2018). Politics of engagement in an age of differing voices. *European Journal of Cultural and Political Sociology* 5 (1–2), 1–11.

Maier, D., Waldherr, A., Miltner P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri H. & Adam S. (2018). Applying LDA topic modelling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures.* 12 (2–3): 93–118.

Middelschulte, H. & Kahle, P. (2019). Ökonomisierung der sozialwissenschaftlichen Bildung? Anwendung eines konventionentheoretischen Diktionärs zur massentextanalytischen Untersuchung einer bildungspolitischen Debatte. In C. Imdorf, R. J. Leemann & P. Gonon (eds.), *Bildung und Konventionen. Die „Economie des conventions" in der Bildungsforschung*. Wiesbaden: Springer, 259–284.

Minsky, M. (1975). A framework for representing knowledge. In P. Winston (ed.), *The psychology of computer vision*. New York: McGraw-Hill, 211–277.

Pearson, K. (1901). On lines and planes of closest fit to a system of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 6 (2), 559–572.

Scharkow, M. (2011). Zur Verknüpfung manueller und automatischer Inhaltsanalyse durch maschinelles Lernen. *Medien & Kommunikationswissenschaft* 59 (4), 545–562.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In E. Tzoukermann & S. Armstrong (eds.), *From texts to tags: issues in multilingual language analysis.* In *Proceedings of the ACL SIGDAT-Workshop.* March 27th 1995. Dublin: ACL SIGDAT, 47–50.

Sellars, W. (2007). *In the space of reasons. Selected essays of Winfried Sellars*. Cambridge: Harvard University Press.

Ylä-Anttila, T.; Eranti, V. & Kukkonen, A. (2022). Topic modeling for frame analysis: A study of media debates on climate change in India and USA. *Global Media and Communication* 18 (1), 91–112.

Ylä-Anttila, T. & Luhtakallio, E. (2016). Justifications Analysis: understanding moral evaluations in public debates. *Sociological Research Online* 21 (4), 1–15.

Ziem, A. (2008). Frame-Semantik und Diskursanalyse – Skizze einer kognitionswissenschaftlich inspirierten Methode zur Analyse gesellschaftlichen Wissens. In I. H. Warnke & J. Spitzmüller (eds.), *Methoden der Diskurslinguistik. Sprachwissenschaftliche Zugänge zur transtextuellen Ebene*. Berlin, Boston: De Gruyter, 89–116.

Appendix 1. Stop Word List

Ort, letzt, lassen, eigen, erst, weit, verschieden, unv., schwierig, meinen, nehmen, richtig, wichtig, klar, mögen, ganz, Laut, tun, tatsächlich, sollen, denken, Thema, Grad, heißen, sehen, lachen, groß, Beispiel, finden, Ähm, wirklich, glauben, wissen, wollen, genau, gehen, einfach, Organisation, kommen, müssen, ähm, geben, machen, sagen, können, werden, haben, sein

Appendix 2. Results of the unseeded topic modeling, full extent

Topic 1	mensch 'human'/'person', stehen 'to stand', mann 'man', bekommen 'to get', halten 'to hold'/'to stop', leben 'live'/'to life', **aktion 'action'**, stadt 'town'/'city', **schreiben 'to write/writing'**, gespräch 'conversation', bringen 'to bring', lesen 'to read', gehören 'to belong', **politik 'politics'**, passieren 'to happen', euro 'Euro', antrag 'request', moment 'moment', dürfen 'to be allowed', kontakt 'contact'
Topic 2	**leute 'people'**, arbeiten 'to work', gut 'good', landkreis 'county', arbeit 'work', unterkunft 'accommodation', total 'total', sprechen 'to speak', **persönlich 'personal'**, privat 'private', klein 'small', wahrscheinlich 'probable', einzeln 'single'/'separate', schön 'beautiful', politisch 'political', zeit 'time', problem 'problem', unterbringen 'to accommodate', völlig 'totally', schlecht 'bad'
Topic 3	**projekt 'project'**, ehrenamtliche 'voluntary', leute 'people', **sache 'matter/thing'**, jahr 'year', **landkreis 'county'**, gut 'good', **stelle 'office/job'**, geflüchtete 'refugees', arbeiten 'to work', **stadt 'town/city'**, treffen 'meeting/to meet', teil 'part', sitzen 'to sit', zeit 'time', gucken 'to watch', **kollege 'colleague'**, laufen 'to walk/be in process', kennen 'to know', person 'person'
Topic 4	jahr 'year', leute 'people', anfang 'beginning', bündnis 'alliance', mann 'man', klein 'small', arbeit 'work', tisch 'table', kümmern 'to take care', aktiv 'active', **stadt 'town/city'**, **sache 'matter/thing'**, rund 'round', laufen 'walk'/'be in process', veranstaltung 'event', aktion 'action', ungefähr 'approximate', lang 'long', rechts 'right', gucken 'to watch'
Topic 5	leute 'people', **geflüchtete 'refugees'**, arbeit 'work', problem 'problem', freiwillige volunteers', **brauchen 'to need'**, cool 'cool', **sachen 'matters/things'**, helfen 'to help', **deutsch 'German'**, anfang 'beginning', relativ 'relative', voll 'full', begeistern 'to inspire'/'thrill', erfahrung 'experience', hand 'hand', deutschland 'Germany', zahlen 'to pay', freund 'friend', wenden 'to turn'
Topic 6	**schreiben 'to write/writing'**, problem 'problem', versuchen 'to try', schnell 'fast', afd 'AfD' (right-winged political party), widrigkeit 'adversity', **kultur 'culture'**, fallen 'to fall', **abschieben 'to deport'**, **abschiebung 'deportation'**, minute 'minute', falsch 'wrong', brief 'letter', job 'job', weise 'way'/'smart', super 'super', klappen 'be working'/'be functioning', geschichte 'history', kulturell 'cultural', äh 'ehm' (filler/interjection)
Topic 7	**leute 'people'**, ehrenamtlich 'voluntary', projekt 'project', gut 'good', brauchen 'to need', frau 'woman', jahr 'year', mensch 'human'/'person', bereich 'domain'/'area', gruppe 'group', bekommen 'to get', **kontakt 'contact'**, suchen 'to search', spielen 'to play', engagement 'engagement'/'commitment', helfen 'to help', arbeit 'work', **verein 'association'/'club'**, ehrenamt 'volunteering', angebot 'opportunity'/'offer'
Topic 8	ehrenamtlich 'voluntary', ehrenamt 'volunteering', tischen 'tables', rund 'round', **verein 'association'/'club'**, kirche 'church', **treffen 'meeting'/'to meet'**, zahlen 'to pay', **engagieren 'to be committed'/'to be engaged'/'to engage'**, nennen 'to call'/'to name', frau 'woman', schnell 'fast', raum 'room', kriegen 'to receive'/'to get', geld 'money', veranstaltung 'event', einzeln 'single', fehlen 'to lack', sport 'sports', markt 'market'

Topic 9	mensch 'human'/'person', solidarität 'solidarity', helfen 'to help', leben 'to live', gesellschaft 'society', dürfen 'to be allowed', brauchen 'to need', jahr 'year', frau 'woman', schaffen 'to accomplish', **situation 'situation'**, problem 'problem', verstehen 'understand', gleich 'equal'/'soon', fragen 'to ask', hilfe 'help', politisch 'political', versuchen 'to try', solidarisch 'solidary', stehen 'to stand'
Topic 10	geflüchtete 'refugees', erleben 'to experience', toll 'great', gut 'good', mensch 'human'/person', sache 'matter'/'thing', bild 'picture', leute 'people', merken 'to remember', **kultur 'culture'**, gewiss 'certain', reden 'to talk'/'to speak', hören 'to hear', stark 'strong', erfahrung 'experience', alt 'old', betreffen 'to concern', sprechen 'to speak', ja 'yes', bleiben 'to stay'
Topic 11	gut 'good', leute 'people', problem 'problem', frau 'woman', **arbeiten 'to work'**, kriegen 'to receive'/'to get', jobcenter 'employment office', landkreis 'county', mensch 'human'/'person', okay 'okay', fragen 'to ask', passieren 'to happen', stehen 'to stand', uhr 'clock', anfangen 'to begin', jahr 'year', arbeit 'work', persönlich 'personal', tagen 'days', tag 'day'
Topic 12	**stadtteil 'quarter/neighborhood'**, alt 'old', fahrrad 'bike', flüchtling 'refugee', straße 'street', liegen 'to lay', erwarten 'to expect', reparieren 'to repair', vertreten 'to represent'/'to advocate', flughafen 'airport', fallen 'to fall', **kennen 'to know'**, werkstatt 'workshop/garage', ausschuss 'committee', flüchtlingsheim 'refugee shelter', besonderheit 'peculiarity', fahren 'to drive', laufen 'to walk'/'to be in process', fehlen 'to lack', quarantäne 'quarantine'
Topic 13	leute 'people', **deutsch 'German'**, kind 'child', gut 'good', jahr 'year', frau 'woman', familie 'family', mensch 'human'/'person', flüchtling 'refugee', mann 'man', **bekommen 'to receive'/'to get'**, laufen 'to walk'/'to be in process', **sprechen 'to speak'**, **lernen 'to learn'**, kennen 'to know', jung 'young', klein 'small', bringen 'to bring', stadt 'town'/'city', projekt 'project'
Topic 14	gut 'good', person 'person', stück 'piece', bereich 'domain'/'area', **struktur 'structure'**, **stadt 'town'/'city'**, stehen 'to stand', versuchen 'to try', **kommune 'municipality'**, angebot 'opportunity'/'offer', arbeiten 'to work', ende 'end', frage 'question', **kollegin 'female colleague'**, äh 'ehm' (filler, interjection), verein 'association'/'club', teil 'part', ehrenamtlich 'voluntary', raum 'room', bleiben 'to stay'
Topic 15	**project 'project'**, jahr 'year', gut 'good', mensch 'human'/'person', **stellen 'offices'/'jobs'/'to provide'**, arbeit 'work', verein 'association'/'club', netzwerk 'network', fragen 'to ask', arbeiten 'to work', bereich 'domain'/'area', **stadt 'town'/'city'**, neu 'new', integration 'integration', politisch 'political', versuchen 'to try', geflüchtet 'fled', land 'land', fördern 'to support'/'to foster', aktiv 'active'

Appendix 3. Results of the seeded topic modeling, full extent

Politics	schreiben 'to write'/'writing' ✓, situation 'situation' ✓, politik 'politics' ✓, aktion 'action' ✓, politisch 'political' [6], positive 'positive', partei 'political party', halten 'to hold'/'to stop', aktiv 'active', abschiebung 'deportation' ✓, zeigen 'to show', medium 'medium', afd 'AfD' (right-winged political party), rechtlich 'legal'/'judicial', presse 'press', abschieben 'to deport' ✓, bekommen 'to receive'/'to get' [2], stimmen 'votes'/'voices'/'to be correct', hören 'to hear', wählen 'to vote'
Refugees	mensch 'human'/'person', gut 'good', frau 'woman', leute 'people' [4], deutsch 'German' ✓, sprechen 'to speak' ✓, arbeiten 'to work' ✓, problem 'problem', bekommen 'to receive'/'to get' ✓, brauchen 'to need' ✓, stehen 'to stand', jahr 'year', kriegen 'to receive'/'to get', flüchtling 'refugee', fragen 'to ask', mann 'man', arbeit 'work', deutschland 'Germany', geflüchtete 'refugees' ✓, kind 'child'
Organization	projekt 'project' ✓, stadt 'town'/'city' ✓, landkreis 'county' ✓, jahr 'year', stellen 'office'/'jobs'/'to provide' ✓, gut 'good', arbeit 'work', integration 'integration', netzwerk 'network', gucken 'to watch', kommune 'municipality' ✓, ehrenamtlich 'voluntary', person 'person', sache 'matter'/'thing' ✓, struktur 'structure' ✓, Aufgabe 'task', neu 'new', angebot 'opportunity'/'offer', land 'land', versuchen 'to try'
We	leute 'people' ✓, verein 'association'/'club' ✓, gut 'good', mensch 'human'/'person', treffen 'meeting'/'to meet' ✓, jahr 'year', kontakt 'contact' ✓, kennen 'to know' ✓, persönlich 'personal' ✓, engagieren 'to be committed'/'to be engaged'/'to engage' ✓, laufen 'to be working'/'to be functioning' [5], helfen 'to help', neu 'new', gruppe 'group', veranstaltung 'event', ehrenamt 'voluntary work', anfang 'beginning', arbeit 'work', okay 'okay', engagement 'engagement'/'commitment'
Industrial	bereich 'domain'/'area' ✓, good 'gut', zeit 'time' ✓, organisieren 'to organize' ✓, gemeinsam 'together', schaffen 'to accomplish' ✓, jahr 'year', rund 'round' ✓, funktionieren 'to be working'/'to be functioning' ✓, unterstützen 'to support', gemeinde 'community', kirche 'church', unterschiedlich 'different', tisch 'table', bestimmt 'certain', ehrenamtliche 'volunteers', klein 'small', problem 'problem', tischen 'tables', laufen 'to walk'/'to be in process' ✓
Civic	mensch 'human'/'person', solidarität 'solidarity' ✓, politisch 'political' ✓, gesellschaft 'society' ✓, leben 'life'/'to live', solidarisch 'solidary' ✓, fragen 'to ask', gleich 'equal'/'soon' ✓, neu 'new', verstehen 'to understand', verein 'association'/'club' ✓, möglich 'possible', verändern 'to change', versuchen 'to try', stark 'strong', gemeinsam 'together', gesellschaftlich 'social'/'societal', unterstützung 'support', fallen 'to fall', meinung 'opinion'
unseeded Topic 1	wohnung 'apartment', unterbringen 'to accommodate', privat 'private', wohnraum 'living space'/'housing space', problem 'problem', familie 'family', sozial 'social'/'societal', wohnen 'to live'/'to lodge', deutschland 'Germany', n 'n', dezentral 'decentralized', verhindern 'to prevent', kennen 'to know' [4], block '(apartment) block', betreiber 'provider', aufbauen 'to build up', problematisch 'problematic', staat 'state', bezahlbar 'affordable', probleme 'problems'

unseeded Topic 2	mensch 'human'/'person', uhr 'clock', woche 'week', tagen 'days', englisch 'English', stunden 'hours', fahren 'to drive', politisch 'political' [6], karte 'map', arzt 'doctor', sitzen 'to sit', heimat 'home (country)', versorgung 'supply', behandeln 'to treat', tee 'tea', quarantäne 'quarantine', verteilen 'to distribute', routine 'routine', aktuell 'up-to-date'/'present', bus 'bus'
unseeded Topic 3	arbeiten 'to work' [2], sache 'matter'/'thing' [3], jugendliche 'youths', land 'land', beratungsstelle 'information center', frau 'woman', extrem 'extreme', grund 'reason', anspruch 'claim', kennen 'to know' [4], arbeitslosengeld 'unemployment benefit', menschenhandel 'human trafficking', bürgermeister 'mayor', afrika 'Africa', Freund 'friend', opfern 'victims', auszahlen 'to pay off', schritt 'step', bevölkerung 'population', asyl 'asylum'
unseeded Topic 4	arbeit 'work', geflüchtet 'fled', jahr 'year', fragen 'to ask', gruppe 'group', freiwillige 'volunteers', kind 'child', jugendlich 'adolescent', emotional 'emotional', ort 'place', unglaublich 'unbelievable', gelten 'to be valid'/'be in force', erwartung 'expectation', querschnittsthema 'cross-sectional topic', vormund 'legal guardian', stark 'strong', letztendlich 'at long last', erfolgreich 'successful', belastung 'burden', erhalten 'to receive'/'to get'
unseeded Topic 5	kind 'child', familie 'family', sache 'matter'/'thing' [3], mann 'man', reden 'to speak', toll 'great', eltern 'parents', verstehen 'to understand', ja 'yes', tochter 'daughter', interkulturell 'intercultural', lesen 'to read', frühstücken 'to eat breakfast', erzählen 'to tell', kultur 'culture' [1], zuwanderer 'immigrant', mädchen 'girl', erleben 'to experience', einkaufen 'to shop'/'to buy', hintergrund 'background'
unseeded Topic 6	bild 'picture', frau 'woman', ausstellung 'exhibition', passieren 'to happen', mann 'man', schicken 'to send', platz 'space'/'place', zweit 'second'/'in pairs', Schiff 'ship', kopftuch 'headscarf', foto 'photo', fotografieren 'to take a picture', erleben 'to experience', tragen 'to carry', freundlich 'friendly', werfen 'to throw', kapitän 'captain', standbein 'mainstay', großzügig 'generous', polizist 'police man'
unseeded Topic 7	kollegin 'female colleague' [3], tag 'day', büro 'office', kurz 'short', zuständig 'responsible', e-mail 'email', türe 'door', telefon 'telephone', termin 'appointment', frage 'question', zweit 'second'/'in pairs', anrufen 'to call', handlungsfeld 'field of action', warten 'to wait', bitten 'to request', wohnung 'apartment', zeit 'time' [5], schicken 'to send', klient 'client', team 'team'

Political Polarisation on Digital Media: An 'Up Next' Algorithm Analysis of Political Videos on YouTube

Yu-Ning Chuang
University of Sheffield, Department of Sociological Studies
E-mail: ning830717@gmail.com

Abstract
This article uses algorithm analysis to examine how political content is spread and whether it brings out political polarisation. Researchers describe YouTube as an emerging tool for media communication through shifting from one-way to two-way, which brings up a series of discussions of 'digital bias' (Schneider 2018; Billig 2009), especially in political communication. Burgess and Green (2018) mentioned that YouTube has not only increased the scale and complexity of its commercial practices but also the rules to contribute to the platform. More tensions between mainstream culture and subculture have emerged, like political struggles and conflicts of interest. As YouTube increases in popularity as a platform for people engagement, it is concerning that YouTube's recommendations for videos have become an approach to promoting totalitarian ideology (O'Callaghan et al. 2015; Tufekci 2018). In terms of research steps over analysing the 'Up Next' algorithm on YouTube, I deal with each of the video network analyses through two stages of network analysis, including integrated analysis and cluster analysis via Gephi software. By analysing the 'Up Next' algorithm for 473 political talk shows videos with neutral political content, right-wing populist content and left-wing populist content, YouTube leads viewers to political polarisation through its algorithm. Firstly, by further examining the different sizes of nodes, it can be found that these clusters sometimes focus on a certain event in the same period. Secondly, some of the clusters are gathered by the same programme and often have the same political bias, which indicates that users might be directed towards more and more extreme videos through YouTube's recommendation algorithms when they watch certain videos. In this paper, I discuss the role of digital media in political communication, which helps to define whether YouTube is a tool or a medium in political communication.

Keywords: YouTube, political polarisation, algorithm analysis, political talk shows, Up-Next recommendation

1. Introduction

YouTube has become an emerging tool for media communication by shifting from one-way to two-way communication. This shift has sparked a debate on 'digital bias', which is issued by digital platform, such as the composition of the users may be led to imbalanced content. The digital bias highlights the digital sociologists need to clearer identify the relations between digital social research and digital social life, especially in political communication (Schneider 2018; Billig 2009). Marres (2017, 101) pointed out three problems of digital bias: (1) biased data and content, (2) biases built into research instruments, and (3) methodological bias.

Post-truth politics is a kind of political culture which has gained attention in recent years (Davies 2016). In popular use, it is associated with an increasing disregard for factual evidence in political discourse (Lockie 2017), meaning that it has subverted an "actions speak louder than words" approach to prioritise eloquence over actions. The interactive functions of platforms like YouTube have blurred the line between news and social media. In this environment, post-truth politics has become particularly problematic.

Post-truth politics can often cause serious 'political polarization.' As Jasanoff and Simmet (2017) explain, post-truth politics means that facts used in policy are usually for democratic contestation and deliberation. Runciman (2018) emphasized that post-truth politics undermined the basic distinction between reason and emotion in modern politics, in that people put emotions first and evidence and truth second.

The emergence of political talk shows was a notable example of the mediatization of Taiwanese television, that political talk shows can be seen as a field whereby the mass media influence political

society. In the 1980s, transformation especially mentioned the multiple communication techniques, 'neoliberalism' turned the news industry's structure towards 'marketization' and 'privatization'. The two changes affected traditional media interaction, which meant that the dominance of news tended to satisfy the preferences of the audience or business owners in order to weaken the control of the government or ruling party over news broadcasts. Along with the development of new media, political talk shows started to change the form of communication with viewers from 'one-way' communication to 'two-way' communication. Similarly, a modern equivalent is the five functions on YouTube for online communication: recommendations, comments, channels, uploads and views.

According to Burgess and Green (2018), YouTube not only increased the scale and complexity of its commercial practices but its control over participation in the platform. The multiple functions of YouTube keep the independence of commercial development and ensure the reality, and both mainstream culture and subculture are published on the platform. As tensions between mainstream culture and subculture emerge, such as political struggles and conflicts of interest, it is evident that YouTube is a popular platform for people engagement. Existing literature suggests that the political content on the recommendation system shows how YouTube's recommendations algorithm is becoming a gateway to accessing extremist ideologies (O'Callaghan et al. 2015; Tufekci 2018).

This article analyses the 'Up Next' algorithm through integrated and cluster analysis, to examine how political content is spread and whether it brings about political polarisation.

2. Power Change and the Political Effects of Television

This chapter contains two sections: The first section illustrates the development of television during the post-World War II period and points out the shifting of communication from one-way (Web1.0) to two-way (Web 2.0). The next section focuses on the formation of mainstream political consciousness in Taiwan and the form of the pan-blue populists (the Kuomintang (KMT)) and pan-green populists (the Democratic Progressive Party (DPP)), which primarily lead political consciousness in Taiwan.

2.1. The Development of Taiwanese Political Talk Shows

The development of television in Taiwan is linked to a shift in political power, which can be categorised into three time periods: (1) 1945 – 1997: the hegemony of the government (the KMT) controlled post-World War II, where television was a tool for government propaganda and a material symbol of innovation, as it was in most leading countries of the time; (2) 1997 – 2006: when the media became regarded as the fourth pillar of the state through the political democracy, as independent of the government, when the agenda-setting of political news focused on public or commercial interests; (3) 2006 onwards: the form of communication changes from 'one-way' to 'two-way' with the development of new media, weakening both government power and its role as the fourth estate. The public began to decide the kind of content they want to watch. These three distinct phases in media responsibility show the degree of political involvement in the media from strong to weak to populist.

The development of television in Taiwan can be traced back to the leading countries of the post-war period, such as the U.S. and Japan. It represented the technical innovation on a material level and the enhancement of the relationship between the media and other disciplines, such as politics, family studies and studies on nationalism. Within this context, the television industry undoubtedly worked in the government's favour at that time, and political propaganda was a part of all discourse. As

Pearson et al. (2014, 14) illustrates, each institution followed acceptable modes of maintaining social operation and continuously changing the relationships and roles of government in society. These power changes have displayed the power of the specific institutions in maintaining dominant ideologies (Pearson et al. 2014, 14). During the post-war period, the ideology took the form of a dominant social value: despotism, which means the television was working in favour of the government, and most content that appeared on TV were controlled or scripted by the government. Since the establishment of the Cable Radio and Television Act in 1993, the number of TV programmes has reached the hundreds and has become widely diversified in theme and ideology (Wang and Chen 2011, 14). However, there are two of main specific political inclination (the KMT and the DPP) which still exists in some of the mainstream media.

Today, the ideology of social value has transferred to individualism, and the development of social media shows how TV ratings are no longer the only index of value for a piece of news. News content providers not only have propaganda-specific political inclinations but also need to consider the like-to-dislike ratio and the comments, namely, consumers' opinions. It forms a specific watching experience in Taiwan, the masses dictate the news being produced most of the time. However, in communicating political issues, the primary purpose of producing politic-based news or programmes is usually to deliver political views. For instance, the Ctitv, a well-known pan-blue TV station, reported substantially Guo-Yu, Han, the KMT presidential candidate at that time, while questioning the legitimacy or capabilities of other presidential candidates. During the Taiwanese presidential election in 2020, the National Communications Commission (abbreviated as NCC) published an investigation, which found that nearly 70% of the Ctitv's airtime was spent reporting on the KMT presidential candidate, which raised the question of media bias and imbalances in news reporting (NCC 2019).

In the second period, when martial law ended in 1987, the media industry's role as a tool of government propaganda and state-approved entertainment ended. At that time, this was not only evident in various types of media, but as the reporting style of 'the old three' started to fail, various controversial reports became mainstream. 'The old three' is derived from the three oldest wireless television stations (TTV, CTV and CTS), established by Taiwan in the 1960s and 1970s, which were controlled by the KMT. In 1993, the government opened a cable TV station, and the first privately-owned TV station was founded in 1997, which meant the law of the Regulations for the Control of Radio, Radio and Television Receivers during the Period of Mobilization had been abolished, and the television was no longer meant only for compulsory government propaganda.

With the emergence of political democracy, the media had a new role: as a watchdog regarded as the fourth estate, which not only played the role of a provider of information necessary for rational debate but also as a mediator between the people and the government. The emergence of the fourth estate is also realized by programme diversity, and as of May 2020, the quantity of programmes has already exceeded 150. The several debates also accompanied by this rising institution, e.g. "Are government funded and government regulated media institutions used for public service or are they propaganda mouthpieces?" and "When private corporations own the media are they furthering their own commercial interests or the public's?" (Dahlgren 1995, quoted in Pearson et al. 2014, 16)

Additionally, on the face of it, after the establishment of the NCC in 2006, the government no longer held control over the media. Simultaneously, the turning point of government control was that it withdrew its official shares from news and media channels. Following this, in the 2008 Taiwan presidential elections, the structure of the news industry was affected by the value of news and business mechanisms rather than by government control (Lin and Lo 2010, 85). At the same time,

political talk shows were developing and becoming increasingly popular and began to occupy the prime time of most TV news channels. The invisible influence of business and consumers continued with the emergence of political talk shows, which became known as mediatization of politics. In general, the concept of 'mediatization' means to the "long-term interrelation processes between media change and social and cultural change" (Hepp, Hjarvard, and Lundby 2010, 223), which means media and society are highly interaction. Mediatization is a complex concept (e.g., Schulz 2004; Krotz 2009), as it is difficult to define and reflects the various influences on media, such as the specificity of different media and how they were used as tools to change culture and society, for better or worse.

However, mediatization is specifically relevant to illustrate the process of technological innovation and institutional change. The emergence of political talk shows was a notable example of mediatization in Taiwanese television. In the 1980s, the transformation especially mentioned the multiple communication techniques and 'neoliberalism' changed the news industry's structure to one of 'marketization' and 'privatization'. These two phenomena affected traditional methods of media interaction, meaning a new dominance of news that sought to satisfy the preferences of the audience and/or business owners and weaken the control of the government or ruling party. The field of political communication and political sociology has continued discourse about television and how its change in structure affected politics (e.g., Holbert 2005; Bolin 2014; Jones & Soderlund 2017). As Bolin states, "The second of these trends hold that this is especially harmful to the relationship between journalists and politicians, as politics and political discourse have been drawn into the entertainment logic of television." (Bolin 2014, 337) Political discourse became focused on entertainment than deliberation and regulation of democracy. However, this study regards entertainment to attract viewers' attention, and the final purpose of political discourse is still for deliberation.

Lastly, along with the development of new media, such as social media, political talk shows started to change the form of communication from 'one-way' communication to 'two-way' communication, observed by the movement of media companies onto social platforms such as YouTube, which provide new ways to interact with content through feedback mechanisms. However, the introduction of multiple types of communication also increased debates in media studies on the bias within media. Pearson et al. (2014, 33) explains, "Media technologies take many forms, ranging from the technological apparatus of …, television to digital technologies associated with the Internet. In each case, the technologies which are used involve a complex network of elements, whose role within the process of mediated communication has been a source of debate and contestation within media and cultural studies." The political talk shows provide a specific example of TV media, and it has become a 'hybrid format' which includes the programme content, public comments on new media platforms such as YouTube and whether the talk shows participants respond to the comments or not. As Baum (2005) points out, the higher the frequency of viewers watching the talk show, the more viewers are willing to participate in political activities. These communications also provide viewers with a sense of political participation. The extent of media coverage on these issues was also positively linked to the public's enthusiasm for such topics (Ader 1995).

2.2. The Formation of Mainstream Political Consciousness in Taiwan

As mentioned above, the political opinions of Taiwan's TV media moved towards liberalization and then polarization after they broke away from the era of the KMT's hegemony over the news. Taiwanese electoral politics have been dominated by two major political parties since shortly after the 2000 Taiwanese presidential election, which is called the first peaceful transition of power. After

the 2000 Taiwanese presidential election, the construction of the mainstream political consciousness in Taiwan has consisted of the pan-blue populists (KMT) and pan-green populists (DPP).

The related studies in Taiwan were focusing on the policy analysis of the television industry and the political economy of the role of media. They supplied detailed analysis and records for drawing up the internal policies, capital structure, and national policy after the period of building the television industry (e.g., Su 1993; Fang 1994; Cheng 2002; Chang 2005; Lin 2005). According to previous studies, the development of Taiwanese television has had an inseparable link with politics. Ko (2008, 118) illustrates how television played a role in political communication, in that he considered the initial role of television as a tool for strategic political communication and used to propagate national ideology. Television was used as a culture machine at that time, was completely obeyed the national pedagogy and control, both in entertainment and political discourse. Since most policies and political ideology in Taiwan are led by the pan-blue populists and pan-green populists, which often represent opposite opinions, it raises concern for the sense of political polarisation in Taiwan. An increasing polarization in the news media and wider political culture has given rise to concerns that the public sphere is being overtaken by 'post-truth politics.' Relevant literature mentions that this sort of news coverage is typically unspecific, conflict rich, negative in tone, and containing attributions of blame (Galtung & Ruge 1965; Semetko & Valkenburg 2000), which is highly useful for those playing the partisan game, in trying to attack other parties while defending their own party (Thesen 2013).

Taiwanese political talk shows are suitable research subjects to examine the behaviour of political polarisation, as they occupy a large proportion of Taiwanese TV content and channels, of which the pan-blue and pan-green account for about half of each. The political communication in talk shows can be seen as an extension of the news, which has a political issue-based setting and detailed background construction. Hsieh (2018, 7) mentioned that Taiwanese political talk shows reflect deliberative democracy in recent years, with the explicit and implicit reconstruction of political deliberation beyond the normative formal decision-making process. Indeed, political talk shows provide a powerful platform for political discourse, through hosts and guests who play multiple roles in explaining and interpreting the topic, and persuading and entertaining the public. These talks shows have has successfully stimulated public engagement in recent years and has caused an influx in the number of politicians forced to step down due to public opinion. Hsieh (2018, 8) further pointed out a problem that almost all political shows have their own political status, as the ideal version of deliberative democracy was never fully materialized in Taiwan. This obvious political wrestling reflected on corporate sponsors, for instance, the DPP government long-funnelled substantial budgetary resources into the pan-green SET (Sanlih Entertainment TV) and FTV (Formosa TV), to their own political ends. Often, political talk shows use inflammatory titles or subtitles to create hotly debated implications and negative impressions.

3. The Practices of Online Political Communication on The YouTube Recommendation System

YouTube's recommendation system is the only function that is produced by an internal mechanism. Since YouTube's recommendation function is difficult to filter, it has always been seen as a complex issue. The increasing number of related works on YouTube's recommendation system (the related video list) has been studied in existing literature. Davidson et al.'s (2010, 294) study on the formulation of YouTube's recommendations system was based on the ranking of a variety of signals for relevance and diversity (Davidson et al. 2010, 294). They found that 207% of YouTube's

recommended pages were based on the 'Most Viewed' page which was higher than the 'Top Favourited' and 'Top Rated' page (Davidson et al., 296). This study illustrated that the performance of recommendation pages follows majority opinions (the 'Most Viewed' page), rather than personalized preferences (the 'Top Favourited'). In the perspective of political communication, the pattern of YouTube recommendation represents a positive effect on political communication by increasing the exposure rate. For example, political parties can expand their exposure through carefully packaged content created about political issues on YouTube. Zhou et al. (2016) demonstrated how to optimize the list of recommended videos and how the 'Related Video' list is used to improve the performance of YouTube's recommendations. This highlights that, after watching the initially selected video, most of the viewers choose to watch a video on the 'Up Next' list.

One ongoing speculation is that YouTube's recommendation becoming an approach to promote totalitarian ideology. O'Callaghan et al. (2015, 474) illustrated that YouTube's recommendations have played an important role in the online strategy and ideological composition of right-wing extremists. The case study (O'Callaghan et al., 2015, 474) also indicated that YouTube's recommendation system has had a strong influence in shaping the extreme right's ideology. For example, Tufekci (2018) mentions that if one viewed videos of Donald Trump rallies on YouTube during the 2016 US presidential election campaign, the recommended videos on YouTube became more and more extreme in ideology and that the content of the 'Up Next' video always gradually deviated from the initial issue.

4. YouTube Recommendation Algorithms

YouTube's recommendation algorithms can be considered as its most influential function, given that YouTube Chief Product Officer Neal Mohan mentioned that over 70% of the time, the platform' AI-driven recommendations determine the content that viewers consume on the platform (Joan 2018).

In recent years, the algorithm has been continuously optimised and is placed under rigorous inspection, for two main reasons: reinforcing scrutiny and increasing user engagement by optimising the algorithm. The scrutiny is carried out by the official YouTube platform, by filtering out videos which mostly satisfy the optimisation, including the removal of violent content, reducing the spread of harmful misinformation, and recommending videos via personalized recommendations. Filtering out violent content and reducing the spread of harmful misinformation are the two typical conditions of YouTube's content optimisation (YouTube n.d.) to strengthen the source credibility of the platform. In addition, according to Hao (2019), since the algorithm is optimized for people's engagement with videos, it tends to provide personal recommendations for videos, creating an addictive experience. Gielen and Rosen (2016) mentioned that 'watch time' is an important metric to promote videos on YouTube, which is a combination of the following items: views, view duration, session starts, upload frequency, session duration and session ends. Gielen (2016) further demonstrated that 'watch time' has direct correlation to a video being 'successful', and that success is identified by reaching viewership equal to or greater than 50% of the subscriber base in the first 30 days. The calculation of this method is by taking the 'views' and multiplying them by the 'average view' duration.

YouTube's personalised recommendation system aim to create ideologically like-minded information spaces, which has been referenced by numerous authors in the literature on this topic (e.g., Rieder, Matamoros-Fernández & Coromina 2018; Röchert, Weitzel and Ross 2020; Hao 2019). Almost all important factors in determining video visibility on YouTube are derived from user behaviours, in that the platform's primary criteria behind generating the Recommended Videos sidebar is by using signals or tags that match videos that user is already watching or by matching their

search history.

Simultaneously, the over-reinforced personalised recommendation function is a potential problem in the process of algorithm optimisation, which Hao (2019) refers to as 'implicit bias'. In the case of political videos, Hao (2019) indicated that this filtering criterion can quickly guide users to videos with extreme content, leading to political polarisation. This mechanism will eventually exclude other viewpoints and leave the most extreme and controversial videos on recommendation lists. Political polarisation is a common phenomenon on social platforms, which is usually considered to be exacerbated by recommender systems. Related studies have indicated that YouTube recommendation systems catalyse the promotion and establishment of specific political ideologies, especially on right-wing populist and neutral political content (e.g., Röchert, Weitzel & Ross 2020). O'Callaghan et al. (2015, 473) mention that the YouTube recommendation system can result in users not being recommended videos or content that clashes with their existing perspectives or beliefs, potentially leading to immersion within an extremist ideological bubble, especially in an environment where media content is based on feedback and 'two-way communication'. Therefore, it is necessary to examine the strength and relation of the link between political neutrality and extremist videos, and whether the videos in recommendation networks ultimately lead to homogeneous or heterogeneous ideologies.

In addition, the aim of the user's consumption behaviour of political news differs between YouTube and traditional news, in that the goal of the YouTube algorithm developers also includes satisfying the optimisation, including the removal of violent content, reducing the spread of harmful misinformation and recommending videos via personalized recommendations. The aim of the traditional news is to balance the presentation of different political viewpoints. The personalized recommendations of the YouTube algorithm bring up an ethical debate of the over-reinforced personalised recommendation, with academics and activists arguing that YouTube has a responsibility as a gatekeeper to extremist content. It means that user behaviour should not be the only criterion for creating recommending videos but should be considered whether these recommended videos also contribute to specific viewpoints, especially in politically charged videos. Jonas Kaiser, an affiliate at the Berkman Klein Centre for Internet & Society, said "YouTube should spend more energy in understanding which actors their algorithms favour and amplify than how to keep users on the platform," (Hao 2019).

According to the previous studies, YouTube's recommendation system focuses on fixing problems by rolling-type correction, and the ultimate target focuses on the balance between increasing users' engagement and the opinion polarisation on the platform. To reach these aims, the above pieces of literature have suggested some important factors in video visibility on YouTube, including user clicks, the time a user spends watching videos, the time a user spends watching recommended videos, same political ideology, same channel, and the most widely subscribed channel.

5. Illustrating Up-next Algorithm Analysis: Researching Online Representations of Political Polarisation

YouTube's recommendation system is the only function that is produced from an internal system called 'Up Next'. As seen in the previous chapter, the review of literature on the topic illustrated that the 'Up Next' videos have been a complex issue due to their focus on user engagement, which makes it difficult to find precise rules for filtering recommendations. These filtering rules in the rankings provided by the recommendation system are based on multiple criteria. There are some patterns to the performance of the 'Up Next' videos, given that 207% of recommended pages are provided based

on the 'Most Viewed' page (Davidson et al. 2010, 296). According to Matamoros-Fernández & Farkas (2021, 238) latest research, channel queries are given a visibility boost by YouTube's recommendation algorithms. This means that some of the specific channels are consistently recommended daily. Viewer counts and channel subscriptions are believed to be the two factors with the most influence on 'Up Next' video recommendations. However, the different levels of influence between these two queries require examination, since channel subscription numbers do boost video visibility and recommendations compared to video view count.

In terms of research steps of analysing the 'Up Next' algorithm on YouTube, I deal with each of the video networks through two stages of network analysis, including integrated analysis and cluster analysis. Firstly, the integrated analysis can illustrate the comprehensive 'Up Next' network on the YouTube page through the distribution of nodes, such as channel titles and view counts, which are demonstrates in the Data Laboratory mode of Gephi software. Next, I then use the 'Modularity Class,' a filtering tool of Gephi, to generate the clustering layout of the 'Up Next' network. Through the cluster analysis, it can be directly recognised to which ideology or recommendation the seed video will be linked to, and whether these seed video are directed towards a specific ideology or issue. Since there is a lack of relevant research currently available, both the research subject and the research field can be seen as a pilot study of algorithm analysis of political communication on YouTube.

5.1. Research Design

In this article, I selected two Taiwanese political talk shows' YouTube channels as the research subjects; the Belle Show (364K subscribers) as a pan-blue political talk show and Zheng JhihdaoLiao 鄭知道了 (348K subscribers) as a pan-green political talk show, both of which have similar data size and subscribers. The classification of the programme behind each political ideology is based on their composition of hosts and guests, whether they mostly have the same political ideology; and whether the political talk show has the same political ideology as the TV company to which it belongs.

In terms of setting the time period, this study uses the data available between January 2022 and March 2022, which is more relevant and updated to the current political environment, as well as fit for the data scale. During this period both programmes broadcasted normally, and the broadcasts were not interrupted due to special circumstances. Next, I generated the video lists via Video List Modules of the YouTube Data Tool (YTDT), and manually filtered out 481 available videos as the seeds using the abovementioned filtering conditions. Apart from the videos which were not published between January and March 2022, I also exclude videos which have invalid links, and those which have been removed by YouTube.

5.2. Generating and Exploring the Video Network Data

To understand the brief video's genre link from the seed video through YouTube's recommendation algorithms, I set additional parameters from the Video Network Module. To generate a good size network, I set 'crawl depth' to 0 on YTDT, which means the network will show us only how YouTube relates the seed videos to other seed videos with no additional recommendations. Through importing the output from YTDT, the integrated video network can be explored by Gephi. To create a readable overall video network and an identifiable cluster network, I used the setting layout of nodes and edges in Gephi.

As the Modularity Report indicates, the Modularity (Q) is 0.574, and the number of communities is 4. Modularity demonstrated that there are dense connections between the nodes of

the video network. The number of Modularity (Q) means each of the clusters has a high correlation. Through manual observation of each of the videos, I found that most of these videos in both right-wing and left-wing political talk shows discuss similar topics at the same time. There are four available clusters that have edges connected after excluding those nodes connected without edges, which indicates how YouTube gathers these videos together. Figure 1 demonstrates the political talks shows' overall video network, which consisted of four clusters (with different colours). Besides, the network demonstrates a variety of political biases, which are not gathered by quantity metrics, such as programme or view count, but are clustered by the time video was published and the channel which can be seen the political status. For example, the green and purple clusters are from the same channel (left-wing political talk shows) but published in the different time periods (see Figure 1).

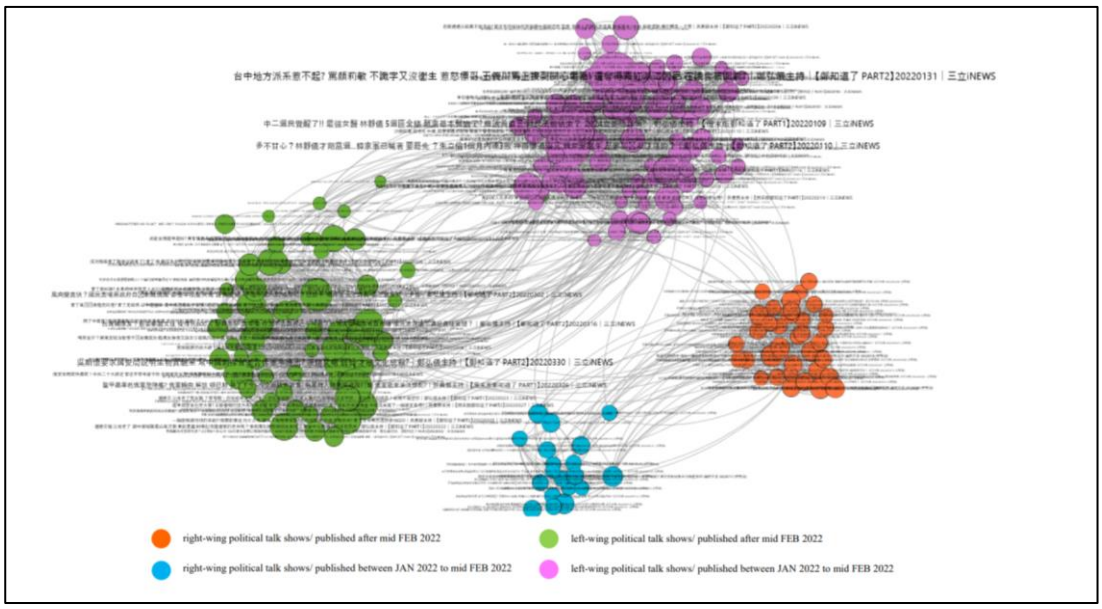


Figure 1: The appearance of the cluster network in Gephi. 2022. [PrintScreen]

5.3 The Issue of Political Polarisation

According to the overall video network analysis, clusters can be identified as smaller groups of seeds and two primarily clustered factors of these seeds are the same channel and the same political status. This has a specific meaning for political talk shows, namely the videos with the same political bias are more likely to appear in the recommended list, which means that videos from the same channel in a cluster will potentially reinforce political polarisation. To investigate the political polarisation caused by YouTube in regards to the content of Taiwanese political talk shows, I set 'crawl depth' to 0 on YTDT, and concentrated on two clusters, right-wing (the KMT) populist content and left-wing (the DPP) populist content. This was used as the starting point to trace the behaviour of the recommendation algorithm. By manually classifying the typical topics of the political talk show videos, Cross-Strait relations between China and Taiwan and criticism of the rival political party are the most frequent topics discussed in these videos. Since the performance of cluster network analysis of the political talk shows does not cluster by topic, I manually filter out two clusters with ten videos as the seeds based on similar topics of the right-wing populist cluster and left-wing populist cluster, including the similar seed size and traffic data, such as view count, like count and comment count.

Based on ten right-wing populist videos as the initial right-wing populist videos seeds, 349 recommendations at crawl 1 were collected. The ten left-wing populist videos were selected and 335

recommendations at crawl 1 were collected. The following table 1 shows the overall initial information on the videos, including both the seeds and the recommendations.

Cluster	Programme genre	Subscribers	Video amounts	Total views	Total likes	Total comments
Belle Show	right-wing political talk shows	364K	59	6,187,541	340,769	8,654
Zheng Jhihdao Liao	left-wing political talk shows	348K	166	15,346,900	180,051	42,390

Table 1: Indicators of the total dataset (including duplicates).

As seen in Figure 2, the yellow dots are the seed videos and the other dots are the recommendations at crawl 1. The squares at the outer ring are the videos to which most videos are eventually directed, the blue square is pan-blue populist videos, green is pan-green populist and black are neutral videos that do not have an obvious political ideology. To get a good network visualisation, I dragged the videos which have a strong direction and connection to other videos to the outer ring and left the seed videos and the videos which do not have a strong direction to the others in the centre. The study has constructed these steps as it is conducive to only focus on the videos which have a strong direction and connection to other videos.

Overall, the recommendation network of the issues of Cross-Strait relations and critiques against rival parties on both pan-blue and pan-green political talk shows have similar effects of communication, that is, the political videos can cause political polarisation through recommendation networks on YouTube. As Figure 2 and Figure 3 demonstrate, the seed videos eventually recommend videos which have the same political ideology. In addition, there are two additional findings from these two figures. Firstly, it can be found that these destination videos are often connected to each other, which means the spread effect of these videos is concentrated on YouTube. It demonstrated that when users browse these destination videos via the automatic recommendation, there will be a high frequency of them being trapped in videos from the same political ideology or ideological bubble. Secondly, almost all videos by a particular channel are from the programmes which I set as the research subject. On further examination of these destination videos, the spread effect of this video genre is relatively closed on YouTube. For example, all recommended videos in Figure 2 are from the same video category (news and politics) and focus on specific political talk shows.

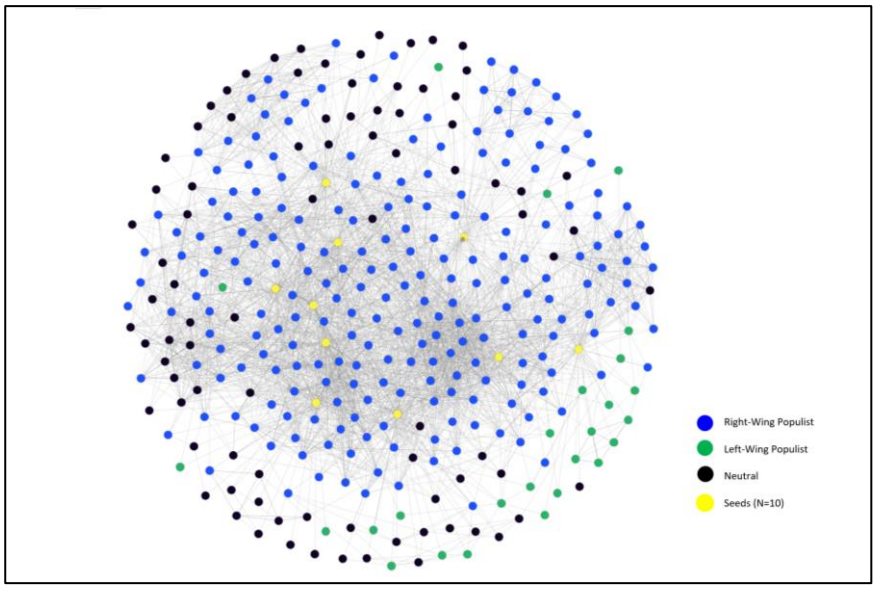


Figure 2: Recommendation network of the right-wing political talk shows cluster. 2022. [PrintScreen]

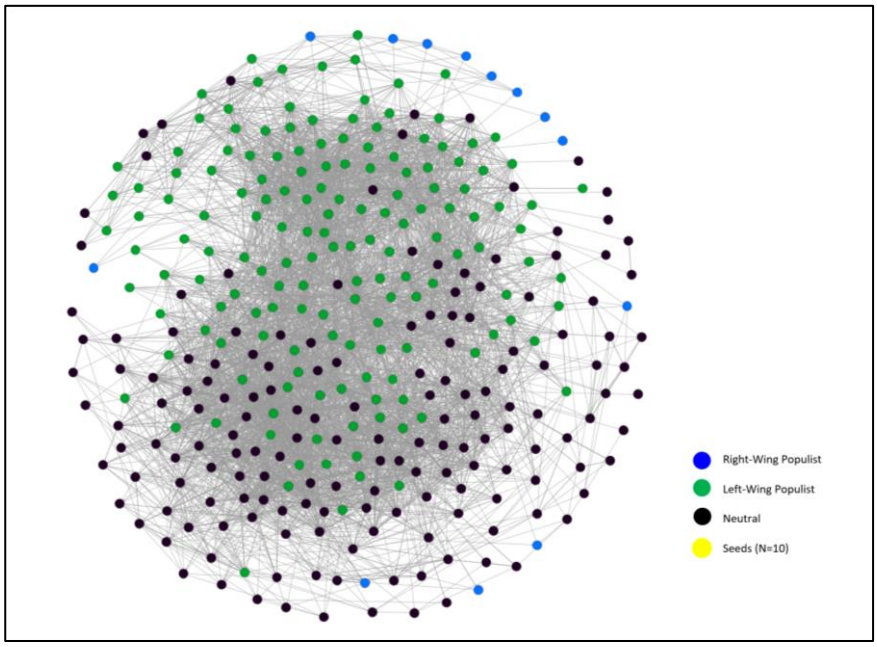


Figure 3: Recommendation network of the left-wing political talk shows cluster. 2022. [PrintScreen]

5.4. Findings from Political Polarisation on Up-next Algorithm Analysis

Overall, it can be found that there is a strong link between the political bias of channels and the YouTube video recommendation system. Based on the selected seed videos, it often leads viewers to other videos by the same channel. In addition, videos which have specific political leanings or personalities are usually recommended for other videos with the same characteristics, which means the YouTube video recommendation system can encourage political polarisation. The political polarisation here is different from totalitarian ideology. Political polarisation here indicates the implicit bias in political viewpoints, and that content consumed confirms theses biases. In terms of political videos, this filtering criterion can quickly guide users to videos with extreme content and lead to political polarisation. From the perspective of specific political issues, such as policy debates,

the filtering algorithm can exclude or limit users' access to videos that represent a different viewpoint than the one they have been consuming or lean towards, and therefore are different from their existing perspective. This can potentially lead to immersion within an extremist ideological bubble. The findings of this study confirm that the political communication on YouTube's recommendation system is becoming a way to receive or promote the specific targeted ideologies (O'Callaghan et al. 2015; Tufekci 2018).

Although there are already many relevant studies on YouTube's recommendation system, there are few YouTube-based studies of Taiwan's political environment. As the relevant studies (e.g., Rieder, Matamoros-Fernández and Coromina 2018; Röchert, Weitzel and Ross 2020; Hao 2019) mention that almost all the important factors in video visibility on YouTube are derived from user behaviours, the primary filtering criteria of generating the recommended videos sidebar is through searching videos that user is watching or matching their search history. Since there is a lack of relevant research currently available, both the research subject (the political talk shows) and the research field (YouTube 'Up Next' system) can be seen as a pilot study of algorithm analysis of the political communication on YouTube.

6. Conclusion

This article has introduced the 'Up Next' algorithm analysis as an approach to examine how political content is spread in social media and whether it brings out political polarisation. First, it is necessary to identify the operation model of YouTube's algorithm for recommendations and that the algorithms are updated at any time. Second, I have mentioned the key role of political videos in the recommendation algorithm analysis, which is currently rarely focused on in Asian studies on political communications. In this article, I have contributed to the recommendation algorithm analysis in social media research online political communication. I have also identified a key issue of the 'Up Next' algorithm analysis that researchers need to address. That is, the algorithm is composed of multiple metrics, including user clicks, the time a user spends watching videos, same political ideology, same channel, and the biggest channel (most subscribers) which makes it difficult to figure out the real control variable for a particular recommended video.

According the study, most of the destination recommendations of the seed videos is often the videos from the same channel and same political status. These demonstrate the typical working behaviour of YouTube's recommendation system is personalised recommendations, which means YouTube recommendations appear to aim to create ideologically like-minded information spaces for their users. The YouTube recommendation system is a double-edged sword for political communication. In the political perspective, this filtering criterion can cause implicit bias by guiding viewers to videos with extreme content (Hao 2019). However, the recommendation system is beneficial for promoting the policy-based issues and creating political awareness so that YouTube users will be recommended videos with similar topics.

Finally, in this article, I have focused on only two political channels, but there are still many others which function similarly. It brings out a methodological challenge in terms of data collection and analysis, particularly at scale. Besides, YouTube's recommendation system is complex and focuses on fixing problems by rolling-type corrections, it is hard to identify the precise factors which determine whether the video appears on a recommendation list.

References

Ader, C.R. (1995). A Longitudinal Study of Agenda Setting for the Issue of Environmental Pollution. *Journalism & Mass Communication Quarterly* 72(2), 300–311.

Baum, M.A. (2005). Talking the vote: Why presidential candidates hit the talk show circuit. *American Journal of Political Science* 49(2), 213-234.

Billig, M. (2009). Reflecting on a critical engagement with banal nationalism–reply to Skey. *The Sociological Review* 57(2), 347–352.

Bolin, G. (2014). Television journalism, politics, and entertainment: Power and autonomy in the field of television journalism. *Television & New Media* 15(4), 336-349.

Burgess, J. & Green, J. (2018). *YouTube: online video and participatory culture Second*., Cambridge, UK; Medford, MA, USA: Polity.

Chang, S. C. (2005). The Commodification of TV Program Production Process in Taiwan. *Chinese Journal of Communication Research* 7, 137–81.

Cheng, Z. M. (2002). Taiwan Dianshih Jhengtse Sihshih Nian De Hueigu Yanjiou Baogao 台灣電視政策四十年的回顧研究報告 [A 40-year retrospective report on Taiwan 's television policy]. *Chuanbo yanjiou jiansyun 傳播研究簡訊 [Communication Research Newsletter]* 31, 1–5.

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B. & Sampath, D. (2010). The YouTube video recommendation system. In Proceedings of the fourth ACM conference on Recommender systems. Barcelona Spain, 26–30 September 2010, 293–296.

Davies, W. (2016). The age of post-truth politics. *The New York Times* 24.

Dahlgren, P. (1995). *Television and the Public Sphere*. London: Sage.

Fang, C. S. (1994). 'Opening-up' of Television Spectrums in Taiwan: A Political Economic Perspective. *Taiwan: A Radical Quarterly In Social Studies* 16, 79–118.

Galtung, J. & Ruge, M. (1965). The Structure of Foreign News. *Journal of Peace Research* 1, 64–91.

Gielen, M. (2016) *WTF Is Watch Time?! Or How I Learned To Stop Worrying And Love The YouTube Algorithm.* https://www.tubefilter.com/2016/05/12/youtube-watch-time-metric-algorithm-statistics/ (Accessed: 14 April 2022).

Gielen, M. & Rosen, J. (2016) *Reverse Engineering The YouTube Algorithm: Part I*. https://www.tubefilter.com/2016/06/23/reverse-engineering-youtube-algorithm/ (Accessed: 14 April 2022).

Hao, K. (2019) *MIT Technology Review*. https://www.technologyreview.com/2019/09/27/132829/youtube-algorithm-gets-more-addictive/ (Accessed: 12 March 2022).

Hepp, A., Hjarvard, S. & Lundby, K. (2010). Mediatization – Empirical perspectives: An introduction to a special issue. *Communications* 35(3), 223–228.

Hsieh, S. C., 2018. *Political Talk Shows in Taiwan: First-and Third-Person Effects, Their Attitudinal Antecedents and Consequences.* University of South Florida.

Holbert, R.L. (2005). A typology for the study of entertainment television and politics. *American Behavioral Scientist* 49(3), 436–453.

Jones, P. & Soderlund, G. (2017). The Conspiratorial Mode in American Television: Politics, Public Relations, and Journalism in House of Cards and Scandal. *American Quarterly* 69(4), 833–856.

Jasanoff, S. & Simmet, H.R. (2017). No funeral bells: Public reason in a 'post-truth' age. *Social studies of science* 47(5), 751–770.

Joan, E. S. (2018) *CNET*. https://www.cnet.com/tech/services-and-software/youtube-ces-2018-neal-mohan/ (Accessed: 12 March 2022).

Ko, Y. F. (2008). The Politics and Discourse of Television: The Formation Process of Television in the 1960s Taiwan. *Taiwan: A Radical Quarterly in Social Studies* 69(3), 107–138.

Krotz, F. (2009). Mediatization: A Concept with Which to Grasp Media and Societal Change. In: K. Lundby, ed. *Mediatization: Concept, Changes, Consequences*. New York: Peter Lang. 19–38.

Lin, L. Y. (2005). Weichyuanjhuyi guojia Yu dianshih: Taiwan Yu nanhan Jhih bijiao 威權主義國家與電視：台灣與南韓之比較 [Authoritarian States and Television: A Comparative Study between Taiwan and South Korea]. *Mass Communication Research*,85, 1–30.

Lin, Y. C. & Lo, V. H. (2010). Taiwandianshihgongsih sihjie zongtongsyuanjyu sinwun baodao jhengdang piancha yanjiou 臺灣電視公司四屆總統選舉新聞報導政黨偏差研究 [Partisan Bias in Taiwan Television Enterprise's Coverage of the Four Presidential Elections in Taiwan: 1996-2008]. *Journal of Electoral Studies*,17(1), 55–90.

Lockie, S. (2017). Post-truth politics and the social sciences. *Environmental Sociology* 3(1), 1–5.

Marres, N. (2017). Digital sociology: the reinvention of social research. Cambridge, England; Malden, Masshuttes: Polity.

Matamoros-Fernández, A., & Farkas, J., 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & New Media* 22(2), 205–224.
NCC. (2019). *2019 Nian 5 Yue Dianshih Sinwun Baodao Guance Baogaoan 2019 年 5 月電視新聞報導觀測報告案 [Television News Report Observation Report in May 2019]*. [Viewed 16 July 2020]. https://www.ncc.gov.tw/chinese/files/19071/8_41689_190717_1.pdf
O'Callaghan, D., Greene, D., Conway, M., Carthy, J. & Cunningham, P. (2015). Down the (white) rabbit hole: The extreme right and online recommender systems. *Social Science Computer Review*, 33(4). 459–478.
Pearson, E., Taffel, S., Nicholls, B., Wengenmeir, M., Chin, K. W., Phillips, H., ... & Urbano, M. (2014). *Media Studies 101*. The Media Text Hack Group.
Rieder, B., Matamoros-Fernández, A. & Coromina, Ò. (2018). 'From ranking algorithms to 'ranking cultures': Investigating the modulation of visibility in YouTube search results', *Convergence (London, England)* 24(1), 50–68. doi:10.1177/1354856517736982.
Röchert, D., Weitzel, M. & Ross, B. (2020). 'The homogeneity of right-wing populist and radical content in YouTube recommendations', in *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, 245–254. doi:10.1145/3400806.3400835.
Runciman, D. (2018). Nervous States by William Davies review – have we really had enough of experts? [online]. *The Guardian*. [Viewed 27 February 2020]. https://www.theguardian.com/books/2018/nov/24/nervous-states-how-feeling-took-over-the-world-william-davies-review
Schneider, F. (2018). *China's digital nationalism*. New York: Oxford University Press.
Semetko, H. & Valkenburg, P. (2000). Framing European politics: a content analysis of press and television news. *Journal of Communication* 50(2), 93–109.
Su, H. (1993). Yuyan(guo/fang) zhengce xingtai 語言（國/方）政策型態 [The National Language and Dialect Policy]. In Cheng Jui-cheng 鄭瑞城 ed. *Jiegou guandian meiti 解構廣電媒體 [Analyzing the Boardcasting Media System]*. Taipei: Chengshe 澄社 (Taipei Society), 217–278.
Thesen, G. (2013). When good news is scarce and bad news is good: Government responsibilities and opposition possibilities in political agenda-setting. *European Journal of Political Research* 52(3), 364–389.
Tufekci, Z. (2018). *YouTube, the Great Radicalizer*. The New York Times [online]. 10 March. [Viewed 10 June 2020]. Avaliable from: https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html
Wang, Y. W. & Chen, B. L. (2011). *Dianshih Meitijhihbo Sinwun Wunzejijhih Yanjiou 電視媒體製播新聞問責機制研究 [Research on Accountability Mechanism of TV Media Production and Broadcasting News]*. Taipei: National Communications Commission. [Viewed 16 July 2020]. https://www.ncc.gov.tw/chinese/files/11121/2732_22837_111219_1.pdf
YouTube (n.d.) *Ever wonder how YouTube works?* https://www.youtube.com/howyoutubeworks/ (Accessed: 12 March 2022).
Zhou, R. et al. (2016) 'How YouTube videos are discovered and its impact on video views', *Multimedia tools and applications* 75(10), 6035–6058. doi:10.1007/s11042-015-3206-0.

Computer-Assisted Investigation of Deixis and Code Switching in Simulated Physician-Patient Interactions

Dániel Mány[1], Andrea Barta[1], Rita Kránicz[2], Renáta Halász[2], Anikó Hambuch[2], Judit Császár[1], Katalin Fogarasi[1]

[1]Semmelweis University, Department of Languages for Specific Purposes

[2]University of Pécs, Department of Languages for Biomedical Purposes and Communication

E-mail: many.daniel@semmelweis-univ.hu

Abstract

In the framework of the course entitled *Professional Language Competencies of Patient-Centered Diagnosis Disclosure* designed for Hungarian medical students (Fogarasi et al. 2020; Halász, Kránicz & Hambuch 2021), teaching assistant students assuming the role of simulated patients and senior medical students playing the role of general practitioners act out medical consultations based on authentic and original clinical documentation (Fogarasi 2018). With a patient-centered theoretical background (Levenstein et al. 1986; Bigi 2016), the aim of the present study is to analyze the use and the occurrence of discourse phrasemes, person deixis and code-switching, as well as their pragmatic and socio-cultural functions and effects on physician-patient interactions (Tátrai 2011; Wagner, Gadebusch-Bondio & Kinnebrock 2020). For this purpose, seventeen interactions were digitally recorded and transcribed using ALRITE speech-to-text software with manual correction afterwards, and then imported to the Sketch Engine corpus analysis tool. The corpus consists of 29,581 tokens and 1040 sentences. Digital functions such as wordlist, word-sketch, key-word extraction, n-gram, concordance, and the visualization function were used to reveal deictic and code-switching elements. A great number of code-switching and deictic elements influencing patients' satisfaction were identified in the corpus and have been implemented into our teaching materials to raise awareness of the importance of their deliberate use. Digital software analysis made it possible to compare written medical reports to related oral consultations with regards to terminology, collocations and communicative strategies.

Keywords: patient-centered communication, diagnosis disclosure, code-switching, person deixis

1. Introduction

Healthcare professionals' and especially physicians' communicative strategies play a major role in patient compliance and health behavior. Patients judge physicians essentially on the basis of their communication; effective communication determines patient satisfaction, improves patient cooperation in therapy, and even contributes significantly to patients' health status. Physicians can effectively help patients make their preferences explicit and then co-construct with them informed preferences to help them reach their therapeutic goals (Bigi 2014. In the present study, while taking into account the aims of patient-centered communication, the aim is to analyze how code switching and person deixis affect physician-patient interactions. The analysis is based on a corpus compiled from written medical documents and transcriptions of interactions between simulated patients played by teaching assistant students and simulated general practitioners (GP) played by medical students in the framework of a course aiming at effective healthcare communication. Both code-switching and person deixes may considerably contribute to the effective transfer of information from physician to patient as well as to the involvement of patients into decision-making regarding their health. Therefore, the results of the present case study may help instructors to further analyze patient-centered communicative strategies in more detail as well as helping medical students to better understand the importance of diagnoses disclosure from a linguistic point of view.

2. Patient-centered communication

Patient-centered communication provides the framework of the present case study. The previous paradigm of approach to medical care was the disease-centered model (Engel 1980). Consequently,

in this model, the physician was the “regulator of the sick condition” (Bigi 2016). The idea to place the patient in the focus of physician-patient encounters first came from Balint (1957); the phrase “patient-centered communication” itself originates from Levenstein et al. (1986). The latter is also the basis for the shared decision making model, which has also been frequently researched recently in connection with the role of code-switching (Fogarasi et al. 2020) and deixis (Kuna 2016) in healthcare communication. Patient treatment fidelity increases in proportion to the degree to which the physician involves the patient in the decision. A shared mind between the participants of the communication can only be realized if the patient understands the physician. By contrast, there is often a mismatch between a physician's level of communication and a patient’s level of comprehension (Weiss 2007). Patients play a crucial role in their healing process, and their compliance influences the treatment’s efficacy. The physician-patient encounter is burdened by the fact that physicians often use their own special language, which the patient is not able to understand. Physicians have often been described as ‘bilinguals’, as they are fluent speakers of everyday language and also of medical language (Bourhis, Roth & MacQueen 1989; Williams & Ogden 2004). Given that the basic ratio should be encouraged to adapt the patient language as far as they code their own medical language (Pendleton 1984), there might be a huge gap in medical education in Hungary, as medical students use their own language for 12 semesters of the curriculum, sometimes between each other and interprofessionally, and find it difficult due to a lack of tools to ‘translate’ it back into the patient’s language. Therefore, implementing code-switching and deictic elements into the discourse can highly contribute to the better understanding of the information by the patient as well as to the involvement of patients into a shared decision-making regarding their health and therapy.

2.1. Code-switching

In medical discourse, code-switching is considered to be a tool for establishing a rapport and convincing the patient. By definition, code-switching takes place when a physician conveys information to the patient using a language which is understandable for the patient. If the physician speaks in their patient’s language or vocabulary, the patient’s comprehension of their illness or condition might increase (Wood 2019). Due to the extensive use of professional terminology in medical documentation, it is essential that patients are provided with detailed explanations of their clinical findings by their general practitioners, so that they can give their consent based on real understanding (Fogarasi et al. 2020). If the physician accommodates the manner of their patient’s speaking, the latter will tend to have more faith in both the diagnoses and the treatment plans. As a consequence, enhanced rapport and trust in the physician may lead to better patient adherence, which is essential in quality patient care (Wood 2019). There is also evidence that the level of health literacy (the ability to obtain, process and understand health information) influences medical care and impairs health outcomes in populations with chronic diseases (Wright 2017).

In physician-patient interactions, the disclosure of the diagnoses plays a prominent role from both a legal and a communicative point of view. Accurate understanding of diagnoses on the part of the patient is an important requirement from a criminal and civil legal perspective (Jobbágyi 2013; Pramann 2017) that must be met by medical communication, and from the perspective of communication strategy (Fogarasi et al. 2020). A key point to the compliance and understanding is partnership between the physician and the patient as well as the appropriate patient education, which is based on the capacity to seek, understand and act on health information (Paterick et al. 2017). According to the American Academy of Family Physicians (2000), patient education is the process of influencing behavior and implementing changes in patients’ knowledge, attitude and skills to

improve their condition. Providing patients with full and relevant information can result in trust between the physician and the patient that empowers patients to participate in their own health care. To achieve this, the disclosure of diagnoses and the physician's verbal explanation of planned measurements and risks should be carried out in an individualized form and the physician must also adapt to the patient's educational background, physical, mental, and psychic condition (Pramann 2017) as severe pain, dementia or panic, e.g., can impair the patient's perception.

Since the precise and unambiguous nomination of diseases in many cases is only possible using Greek-Latin terminology, most physicians find it particularly difficult to change the code to a generally understandable language while disclosing diagnoses, as a pilot study conducted among family physicians in Hungary demonstrated (Fogarasi et al. 2020). Not only did code switching in connection with diagnoses prove inconsistent due to the use of numerous hyperonyms, but the diagnosis disclosure itself was missing as a structural unit in the physician-patient conversation. Most often, patients received information about their diagnoses piecemeal as part of the explanation of therapy; it was not uncommon for them to have to ask follow-up questions to be aware of the rationale for therapy selection or medication dosage (Fogarasi et al. 2020).

2.2. Deixis

Each language displays a great number of words and expressions whose reference can only be interpreted in the light of the circumstances of the utterance. By definition, a deixis is known when the referring expression points to the referent in the context. Deictic expressions are known as contextual elements, given that the interlocutors shall share the same particular context, therefore the meaning of these expressions can only be understood from the contextual elements that are not always displayed in the discourse itself (Safwat 2018). Deictic elements are considered to be the most direct linguistic mirroring regarding the relationship between language and context (Levinson 1983), they thus reflect the relationship between the interlocutors, characterized by the given speech act and sociocultural background. As the present study has a special interest in physician-patient relationship, it takes into account person deixis. It should be mentioned that person deixis is applied to personal relations, by pronouns and grammatical markers in general, while attitude deixis refers to the social relationship, by means of T/V (formal and informal personal pronouns or verbal inflections), forms of address and calls. Person deixis is in connection with the social features of that situation and is generally associated with the roles assumed by the interlocutors (Verschuren 1999; Tátrai 2011; Kuna 2016). As deictic elements are correlated to the linguistic activity as well as to the given context, they represent a perspective (Tátrai 2011), and thus a socio-centric or an egocentric organization in healthcare communication (Domonkosi 2016). Ideally, a physician-patient interaction is characterized by a dynamic peer-to-peer relationship. Nevertheless, physicians' linguistic behavior is more decisive in the interaction, to which code-switching and deixis may highly contribute.

3. Research questions of the present case study

Different structural units of medical genres (e.g., patient personal information, patients' medical history, status, present complaints etc.) are usually marked by typical discourse phrasemes which are considered to be conventionalized formulations, usually based on a noun-verb phraseme (Gréciano 2006). As previous analysis of real GP-patient interactions (Fogarasi et al. 2020) revealed that such phrasemes and the disclosure of diagnoses as an independent structural unit were missing, the present case study examined the use of such phrasemes in interactions led by medical students having been

instructed to dedicate a separate structural unit of the interaction to the diagnosis disclosure.

Besides this, the present case study focused on the use of person deixes, especially, because students were not instructed on their use. They were given the task of establishing a personal, trustworthy relationship with the patient, and the digital analysis of the interactions allowed us to uncover the linguistic means by which this happens spontaneously.

The third research question of our case study was the extent to which medical code-switching took place, i.e., whether the students were able to convey mostly complicated Greek-Latin and English terms, abbreviations, and acronyms in a code understandable to the patient whose profile was enrolled.

4. Data and methods

The corpus of the present case study is compiled from medical reports, and discourses recorded during the course entitled *Professional Language Competencies of Patient-Centered Diagnosis Disclosure* held in the spring semester 2020/21 via the Zoom video chat application as a cooperation between the Department of Languages for Specific Purposes and the Department of Behavioral Sciences of Semmelweis University, Hungary as well as the Department of Languages for Biomedical Purposes and Communication of the University of Pécs, Hungary. The lecturers represented different areas of expertise: Communication Science, Psychology, and Terminology. The accuracy of medical information in the authentic documentation serving as the basis for the re-enacted scenarios was checked by a physician.

4.1. Data

Digitally transcripted texts of the recorded physician-patient interactions carried out by 17 students were included in the present study (12 students in the 6th, 3 students in the 8th, 1 student in the 10th and 1 in the 12th semester of their studies) as well as authentic, anonymized clinical medical reports from four different specialities. The aim of the data collection and the investigation is to make students aware of terminological code-switching and the role of deixes at the level of interaction, and to provide students with effective communicative strategies when giving information to patients. The corpus of the transcribed audio texts consists of 29.581 tokens and 1040 sentences; the corpus of the medical reports consists of 4345 tokens and 230 sentences. The results of the present case study can help formulate possible research questions for a later, statistical-based digital analysis of a larger corpus.

4.1.1. Course structure

The course analyzed in the present study was provided in blocks. First, participants attended introductory lectures on the theoretical background, demonstrating the terminological, psychological and interactional, legal aspects of diagnosis disclosure, as well as the linguistic aspects of subjective disease theories. Following the lectures, students practiced code-switching with the help of a terminology specialist. In the subsequent small-group practical sessions, four students, one teaching assistant student playing the role of the patient, and one instructor worked together. In each group there were four role-plays, with each student having an individual discussion with the 'patient'. After the role-plays, the students received personalized feedback.

4.1.2. The structure of the role-plays

The situational framework for the discussions was that the patient came to the GP's office with a discharge summary from the hospital following his/her hospital stay. As the simulated patient might not have understood everything in the clinic and also because the GP taking over the treatment is obliged to inform the patient again, the patient "brought" his/her medical report and "handed it over" to his/her GP. So, the course being held on-line, both the student playing the role of the GP and the simulated patient students were shown a copy of the same findings.

The medical students attending the course and playing the role of the GP conducted the consultation on the basis of a medical report which had previously been interpreted and explained by a physician to all participants. This step was necessary to make sure medical students have all the relevant medical information, especially when it comes to abbreviations and acronyms, as well as medical terms used only in one medical speciality. The most important communicative goal was to disclose the patient's diagnoses with the necessary explanations, and to answer the patient's questions using the appropriate language code.

The role-plays were 7-10 minutes long and were carried out in the breakout rooms of the Zoom application. As in one round four role-plays were acted out based on medical reports issued by four medical specialties, four medical students, a simulated patient student and an instructor were assigned to a breakout room. After each role-play, a detailed feedback session took place.

Each medical student acting out the GP's role had to lead a physician-patient interaction on one medical report. The medical reports were chosen by a medical doctor for the purpose of this course and represented the most common disciplines of internal medicine: cardiology, pulmonology, hepatology, and oncology, which involves additional terminological and communicative skills needed for breaking bad news. As it was a two-day course, the students had the possibility to act out four role-plays, so each student was able to gain experience in each field. For the purpose of the present case study, one practical unit was recorded parallelly with the consent of all participants in 4 different breakout rooms, including one performance of each medical student as a GP.

4.2. Methods

The Hungarian students' speech production during the physician-patient conversation was recorded in breakout rooms in an anonymized form using the Zoom application as audio files. Afterwards, the audio files were digitally transcribed using the Alrite speech-to-text software. The transcripts were proofread by the participants of the research group and compared with the audio file. In the case of some medical terms and non-confidential proper names (hospitals, clinics, eponyms), manual correction was needed.

Then, the transcriptions of the recorded interactions were assigned to the four medical specialities and a corpus comprising both the medical reports and the transcribed interactions were compiled in Sketch Engine software. The corpus was divided into four sub-corpora, according to the four medical specialties the medical reports represented: Cardiology, Gastroenterology, Hepatology and Oncology. Each sub-corpus was further divided into sub-subcorpora: one containing the written medical report and the other one comprising the transcriptions of the recorded interactions which took place about the respective report. To reveal the weighting chosen by the students playing the GPs, the keywords in the reports and afterwards also the keywords in the transcribed interactions were extracted and compared for each sub-corpus. Also, in each sub-subcorpus containing the interactions on the same report, the word frequency lists were analyzed and the automatized extraction of verbal

phrasemes including n-grams, bigrams (n2) and trigrams (n3) was carried out to investigate possible discourse phrasemes introducing diagnosis disclosure. The most relevant collocations were observed using the visualization function of Sketch Engine. The most frequent types of code-switching were examined based on a frequency list extraction and typical examples were collected from the concordance analysis in the individual sub-subcorpora.

To examine the use of personal deixes, an automatized extraction of first person singular and plural inflectional suffixes (e.g., **ünk, *unk*) and pronouns (e.g., *én* 'I', *ön* 'formal you') was carried out in a separate sub-corpus including all transcriptions of all above-mentioned sub-corpora.

To analyze the extent to which code-switching was performed successfully, the comparison between the keywords automatically extracted in the medical reports and the related interactions was supplemented with the comparison of the word frequency and concordance lists between the reports and the interactions on the reports in all sub-subcorpora. Using the visualization function allowed for the graphical representation of the distribution of specific word combinations, focusing on terminological and deictic features.

5. Results

A great number of code-switching and deictic elements influencing patients' satisfaction were identified in the corpus and have been implemented into our teaching materials to raise awareness of the importance of their deliberate use. In the following, a quantitative, general overview of the corpus and the individual sub-corpora are given as well as the most frequently used strategies are demonstrated.

5.1. The structural demarcation of diagnosis disclosure and the implementation of code-switching

5.1.1. The structural demarcation of diagnosis disclosure

Contrary to our expectations, no discourse phrasemes demarcating the structural unit of diagnosis disclosure could be detected by the extraction of verbal collocations. The interactions turned out to have been started by the GPs with the question how they could help, and the simulated patients usually directly initiated the topic about their not understanding the diagnoses or the procedures they had gone through. The only verb serving a basis for discourse phraseological units in introductions of diagnosis disclosures was found to be *jelent* 'mean' (38 occurrences). Its most frequent collocations occurring in the corpus are shown in Figure 1.

In Figure 1 the pronominal objects *mi* 'what' and *az* 'that' turned out to be the most frequent collocations with 28 occurrences, which together result in the elliptic Hungarian sentence *Mi az?* 'What is that?', asked by the simulated patients. However, all objects of the verb 'mean' were detected to be nouns that were lay language versions of diagnoses, e.g. *csontritkulás* 'osteoporosis', *(magas) vérnyomás* '(high) blood pressure', *májzsugor* 'liver cirrhosis', *gyulladás* 'inflammation', etc. Another verb which was revealed to be closely connected with the disclosure of the diagnoses and was uttered by the GPs was *ért* 'to understand' (53 occurrences). This one, however, was placed after the discussion of the diagnoses and served to make sure that the patient understood everything and the positive reaction given by the simulated patient. Figure 2 demonstrates the most common collocations of the verb *ért.*

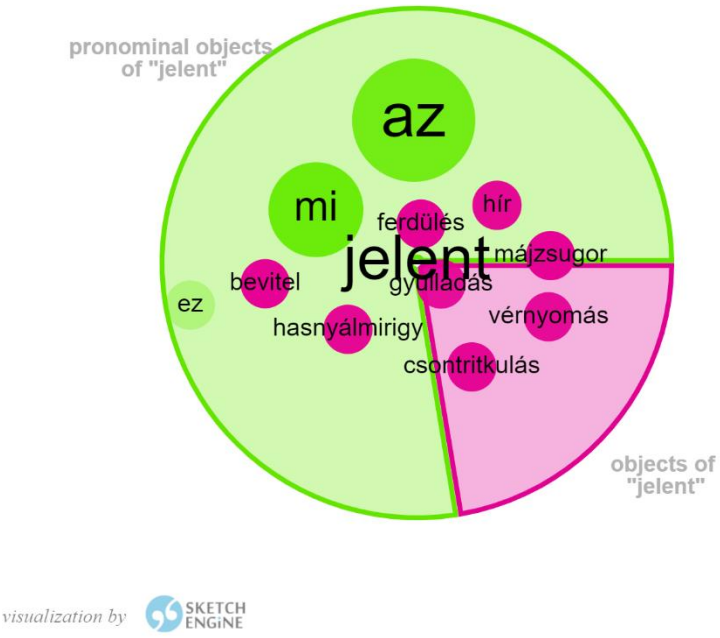


Figure 1: Visualization of the most frequent collocations of the verb jelent ‘to mean’

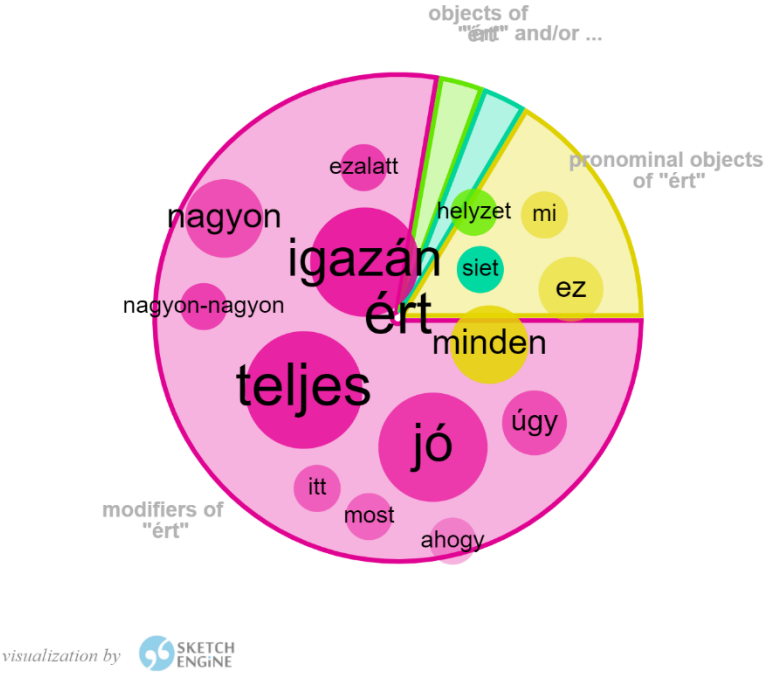


Figure 2: Visualization of the most frequent collocations of the the verb ért ‘to understand’

From the visualization it can be seen that this verb is mostly surrounded by adverbs reflecting the extent of understanding, e.g., *nagyon* ‘very’ (5 occurrences), *nagyon-nagyon* ‘very, very’ (2 occurrences), *teljes* ‘complete’ (7 occurrences), *igazán* ‘really’ (5 occurrences).

5.1.2. Code-switching and deixis

Typically and in our view, code-switching is not a simple word-for-word translation i.e., not simply saying the equivalent of the medical term of Latin-Greek or English origin in Hungarian (as this would usually be just as meaningless to the patient), but rather an explanation or a description of it, using simpler language or paraphrases and adjusting it to the assessed health literacy of the patient.

To reveal the weighting applied by our GP students and the related code-switching strategies, the keywords were digitally extracted in each sub-corpus from both the sub-subcorpora of the medical reports and the ones containing the transcribed interactions. Table 1 summarizes the results of the comparison of the five most frequent keywords in each sub-corpus, named after the specialties they represented.

Table 1 demonstrates that the keywords in the transcribed interactions largely correspond with the keywords found in the medical reports. So, the weighting applied by the impersonated GPs was detected to be highly relevant. However, especially in the interactions referring to the cardiological report, the keywords reflect a rather dynamic approach to the diagnoses. While the medical report focused on the alterations and their locations, in the physician-patient interactions the focus was placed on the prevention of another myocardial infarction by lowering the blood fat content and turning to a dietitian. The same viewpoint can be observed in the interactions on liver cirrhosis, where a liver-friendly diet is recommended. In the other two sub-subcorpora a further step was revealed to be also a main focus: attending a PET-CT for clarification in the case of a malignant tumor and inquiring about possible difficulties with cleaning the stoma bag in the gastroenterology case. To sum up, the results of the keyword comparison carried out in the sub-corpora suggests that code-switching was efficient as the first five keywords mostly corresponded between the sub-subcorpora, and in the sub-subcorpora of the interactions no Latin, Greek or English term occurred.

As the Hungarian language possesses national versions for each medical term of Latin or Greek origin (e.g., *pancreas, stomach, larynx*, or also clinical terms such as *endoscopy* etc.), disclosing diagnoses actually involves a change in terminology. Although patients are usually familiar with the national language variants of medical terms, they are not necessarily aware of their exact medical meanings. Therefore, code-switching does not mean a mere change of words to lay language, but in most cases detailed explanations are needed.

Based on a keyword extraction from the medical reports and the transcribed interactions, the bi- and trigrams terminologically connected to the most frequent keywords were contrasted to identify possible terminological strategies of code-switching in the physician's utterances within the concordance function. As a result of the analysis, four types of code-switching could be differentiated, from a terminological point of view. In Table 2 individual examples of each type are listed in English, with reference to the related medical report.

	Medical report	Related physician-patient interaction
Cardiology	*STEMI* 'ST-elevated myocardial infarction' *RCA* 'right coronary artery' *coronarographia* 'coronarography' *hyperlimidaemia* *tricuspidalis* 'tricuspidal'	*vérzsír* 'blood fat' *befárad* 'to come in' *elzáródik* 'to get clogged' *dietetikus* 'dietitian' *szűkület*, *kitágít* 'narrowing, to widen'
Gastroenterology	*colonoscopia* 'colonoscopy' *colitis ulcerosa* 'ulcerative colitis' *anaemia* *anus praeternaturalis* 'artificial anus' *arryhthmia*	*megoperál* 'to operate on somebody' *vérvesztés* 'loss of blood' *ritmuszavar* 'problem with the regularity of heartbeat' *vesefunkció* 'kidney function' *sztómazsák* 'stoma bag'
Hepatology	*insufficientia* 'insufficiency' *varicositas oesophagi* 'esophageal varices' *portalis hypertensio* 'portal hypertension' *gastropathia* 'gastropathy' *ascites*	*nyelőcső* 'esophagus, gullet' *májkímélő* 'liver-friendly' *elégtelenség* 'failure' *májzsugor* 'shrinking of the liver' *folyadékgyülem* 'accumulation of liquid'
Oncology	*adenocarcinoma* *oncoteam* *nycs (nyirokcsomó)* 'lymph nodes' *irradiation* *metastasis*	*bronchoscopia* *PET-CT* 'Positron Emission Tomography and Computed Tomography' *megnagyobbodás* 'enlargement' *tüdődaganat* 'lung tumor' *hasüreg* 'abdominal cavity'

Table 1: Comparison of the five most frequent keywords retracted from the medical reports and the related transcribed interactions in the sub-corpora, i.e. according to medical specialties

Terminological strategies of code-switching	Medical report	Utterances by the simulated patient	Utterances by the student playing the role of the GP
1. Using a medical term with its explicit explanation	'His complaints are partly due to cardiac decompensation caused by **atrial fibrillation with high ventricular rate**.'	'They said then that there was **some rhythm problem.**'	'Here's where that's written. That is atrial fibrillation. It actually **means that your heart is beating randomly, and that it's beating a bit fast**.'
2. Explanation of a medical term following the simulated patients' enquiries	'The patient was examined in 2016 for **polyneuropathy with sensory axonal focus** – diabetes was suspected, no other etiology was confirmed.'	'It says here that there is **polyneuropathy with sensory axonal focus.** What is that? I might google what that means.'	'I'll find it in a second. Yes. Well, **that's what you're feeling in your leg. The numbness.'**
3. "Hidden" code-switching	'[Diagnosis by] **endoscopy**: multiple gastric and duodenal **ulcers**'	-	'And you had **a tube put down your throat,** I believe. They saw through the tube (they looked down into it) that **there were ulcers. There are small ulcers in your stomach. And they can look at the initial part of the intestine** …'
4. Keeping a medical term after giving a definition	'US and CT confirmed a malformation of the **pancreatic** body'	'I don't know what **pancreas** means.'	'Well, **pancreas means [the Hungarian word for pancreas**]. Do you happen to know what kind of organ it is? (...) We will do our best to treat your pancreas with the most appropriate procedures.'

Table 2. Examples of the four types of code-switching in the transcribed interactions, with references to the medical reports, in English translation.

In Type 1, the explanation was introduced by the phrase 'Here's where that's written' indicating that the GP was looking at the medical document in order to find the medical term that the patient recalled from an oral conversation with his physician. Mentioning the medical terms together with their explanations may help the patient assign the medical term to its meaning and to be able to associate it with his/her disease in future documentation.

Type 2 of code-switching is similar to type 1; however, it is less accurate as the physician did

not repeat the long medical term. He/she just pointed to the symptom that the patient had previously related.

Type 3 was identified in the corpus based on the lack of subordinations and lexical explicitations. Generally, the explanation is introduced by a phrase which refers to the fact that the information was retrieved from a medical document: *ahogy látom* 'as I can see', *úgy láttam* 'I saw that'*, itt írja* 'here it's written', *volt u*gye 'you had a …, didn't you?' *a lelet alapján* 'according to the record'.

Type 4 is the only evidence of returning to the use of a medical term that has been explained beforehand. It shall be noted that even though code-switching is essential in physician-patient interactions, the goal of patient-centered communication is not always to omit the scientific term. This is the reason why strategies that retain scientific terms are also considered as code-switching procedures in the event of the explanation being given by the physician. Often, it is necessary that patients themselves know basic medical terms related to their condition, especially in the case of chronic diseases suggesting a long-term relationship between the interlocutors. As a part of patient education, patients acquire a certain knowledge of their condition, including some basic medical terms, whose use may further contribute to effective communication. In our corpus, the meanings of scientific terms are given by physicians at their first appearance.

Not only code-switching, but also deixis, and especially person deixis plays a major role in physician-patient communication strategies. By definition, a person deixis means the use of expressions to point to a person (Safwat 2018), in Hungarian this is generally done with personal pronouns (e.g., *én* 'I', *nekem* 'for me'), with possessive constructions, markers of the person and number of the possessor (e.g., *diagnózis**unk*** **'our** diagnosis'), with verbal inflectional suffixes marking person and number (e.g., akkor még legközelebb találkoz**unk** 'so **we**'ll meet next time'), definite (e.g., *ez* 'this') and indefinite (e.g., *valaki* 'somebody') pronouns and honorifics [e.g., *doktor úr, doktornő* 'doctor (male, female)'].

Patterns of deixes registered in speech acts all have distinct values when it comes to social relations and metapragmatic functions. The personal pronoun *maga* (formal you) is displayed only two times. The reason that the personal pronoun *maga* is underrepresented in the corpus is that nowadays it is linked to quite an offensive role and no longer has an intimate value. Therefore, the pronoun is rather considered as a feature of top-down communication in hierarchical, asymmetric social relations (Kuna 2016). Consequently, a greater incidence of that pronoun may contradict patient-centered communication, otherwise defined by equality and partnership. In contrast to the small extent of *maga*, the sociocultural value of *ön* (formal, more polite you) results in its frequent use, with 179 occurrences in the corpus, compared to 2 for *maga*. As an addressing element, *ön* is widely used in official and status-marked communicative domains as well as in communication with strangers. In this case, the pronoun represents the interlocutors' adaptation to the official speech situation. While the V (formal voice) form may suggest a distancing value, its use can be suitable for speakers who are not yet on T terms (informal voice). It is also suitable for representing values like politeness and respect, especially in formal communication like physician-patient interactions.

As this version of the V form seems dominant in the corpus, a visualization can reveal some pragmatic functions related to this deictic pronoun, as Figure 3 suggests.

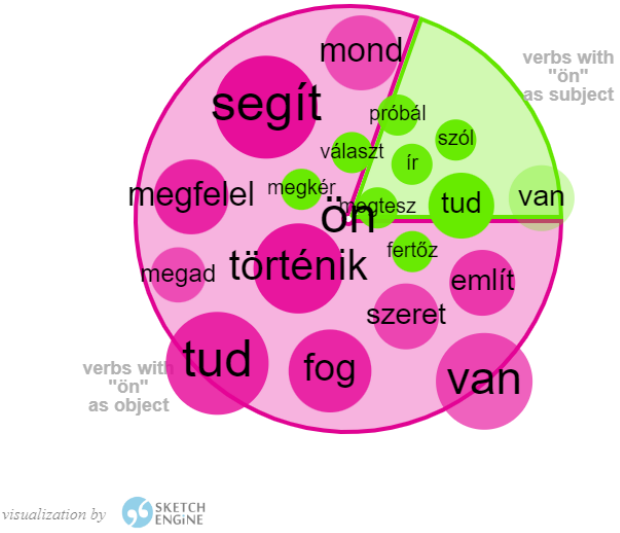


Figure 3: Visualization of the most frequent collocations of the personal pronoun *ön* 'formal you'

As demonstrated in Figure 3, much more verbs were collocated with the pronoun *Ön* having it as an object (93 occurrences) than as a subject (13 occurrences). The other 73 occurrences are displayed in the form of other usage patterns (in accusative, ablative, adessive and alliative). This suggests that the patient was less frequently addressed directly using this pronoun in a subjective function than something related to the patient e.g. the *patient's* symptoms, helping *the patient* or the structure with the verb *van* 'to be' which is used in Hungarian in the sense of 'to have'. This seems to be an alienating use of this pronoun, separating the patient from his or her symptoms and illnesses and maybe making it easier for him or her to look at his or her problem more objectively. At the same time, it might be also easier for the physician to objectify the patient's problem.

Verbs used with the pronoun as an object or a subject might suggest helpfulness (*segít*, 'help'), as in the utterance *szívesen segítek Önnek* ('I am happy to help you'); shared decision-making based on the patient's needs (*választ,* 'choose'; *megfelel,* 'for something to suit somebody') as in the utterances *választ séta és evés mint életmódváltás között* ('to choose between walking and eating as lifestyle changes)' and *megfelel Önnek, ha lépésekben haladunk* ('[something] suits you if we proceed step by step'). They can also convey meanings of politeness (*ker,* 'ask') as in the utterance *kérem, jelezze, ha ezeket a tüneteket tapasztalja* ('I ask you to let me know if you experience these symptoms'), and encouragement and persuasion in connection with patient compliance (*próbál,* 'try'; *csinál* 'to do') as in the utterances *próbáljon figyelni a tüneteire* ('try to pay attention to your symptoms') and *csinál mit csináljon, ha ezeket tapasztalja* ('do what you should do if you experience this'). Finally, they can also refer to taking into account the patient's emotions (*szeret,* 'like') as in the utterance *szeret van még valami, amit szeretne megbeszélni?* ('is there anything you would like to talk about?'), or knowledge related to the patient's health condition (*tud,* 'know') as in the utterance *tudom, hogy vérveszteségem volt* ('I know I lost some blood'). All these pragmatic functions shall further amplify patient-centeredness in connection with functions of deictic elements.

The *tetszik* construction includes the auxiliary *tetszik*, literally meaning '[it] pleases [you]' as well as its infinitival complement, as in *Hogy tetszik lenni?* 'How are you?'(formal), lit. 'How does

it please you to be?', evolved as a marker of politeness in conjunction with pronominal patterns of V (Domonkosi 2018, 70). Generally, this V variant is used among child speakers when they address adults, but also in intimate and personal but not equal social relations, primarily when there is a large age gap between the interlocutors (and the addressee is female), which is often the case when it comes to physician-patient interactions. This tendency can be on the rise in the present corpus, given that medical students are in general under the age of 25, while patients simulated by the instructors usually represent an older generation. Nevertheless, the *tetszik* construction can also be employed among speakers of the same age as an expression of politeness, especially in questions and requests. The *tetszik* construal is considered as respectful but still a direct and warm V variant, therefore it can simplify the speech situation which reflects the status of interlocutors as well as their differences in age and gender (Domonkosi 2018, 71). In the corpus, 21 *tetszik* occurrences can be identified. This V form is linked to an imperative voice, the intensity of which is moderated by the use of *tetszik* in situations when medical students seek to reassure patients or give them advice.

6. Discussion

In our present study, we conducted a terminological comparison of written and transcribed oral texts within the framework of a case study. This comparison was possible because the content of the oral interactions was based on the written texts. As the medical reports were written in a strictly formalized and highly terminologized language, the use of computer-assisted methods made it possible to efficiently identify the key expressions with their verbal references.

The oral interactions could be recorded in parallel with the help of the Zoom application, and transcribed using Alrite software, which is also suitable for the recognition and processing of the Hungarian language. Only a short review was necessary to make small corrections. The written and oral texts were then further analyzed with the help of Sketch Engine software, which supported the terminological-communicative analysis with its numerous functions. The method of corpus analysis allowed subcorpora to be created according to the medical disciplines and sub-subcorpora according to the criterion written or transcribed, but it was also possible to compile new sub-corpora and analyze all transcribed texts in one sub-corpus.

By using the keyword extraction function of the Sketch Engine, it was possible to find the discourse phrasemes around which the diagnosis disclosure was organized. The semantics of these expressions indicate the weighting of the information, which is important from both a communicative and a medical point of view. Based on the extracted keywords, bi-and trigrams contributed to identification of the terminological strategies of code-switching. The examination of the frequency lists allowed for a detailed analysis of the use of personal deixes, extended by a concordance analysis, which made a contextual interpretation possible.

The results yielded by the analyzed corpus suggest that the research questions were relevant but the disclosure of diagnoses was not demarcated by linguistic techniques as an individual structural unit. In this respect, our findings correspond with previous research conducted on real GP-patient interactions (Fogarasi et al. 2020). While in the previous study the written reports could only be compared to the transcripts by manual method, without the possibility of identifying the key expressions of each report and the corresponding interaction, in this analysis, by examining the frequencies of occurrence, it was also possible to identify typical code-switching strategies and to infer the weighting of the information. Since the students impersonating GPs in our course have to interpret the medical report as a whole - similarly to real situations - after reading it once and having to justify their medical decisions, the agreement of the computer-assisted proven weighting of the

written texts and the weighting by the students is also from a medical professional aspect of great importance. Our results also prove the importance of a conscious use of personal deixes, which the students seem to have implemented successfully after having been instructed to be aware of these grammatical and lexicological phenomena.

Despite the evident positive effect of code-switching and person deixes on healthcare interactions, peer-to-peer relationships, and patient compliance, patient-centered communication includes other pragmatic and sociolinguistic constructions, such as empathy, persuasion, politeness, etc., whose empirical research is to be undertaken to obtain a more extensive description of communicative strategies. Also, code-switching and deixes (including space and time deixes) should be further analyzed and categorized in larger corpora. Currently, an investigation of interactions between Hungarian general practitioners and patients is being carried out. It will be clarified whether differences in strategies used by medical students and general practitioners can be identified, and thus implemented in our course program.

7. Conclusion

In our case study, the speech production of medical students was analyzed while acting out diagnosis disclosure with a teaching assistant student in the role of the patient, from the aspects of code switching, deixis and terminology. Four strategies of switching Greek-Latin medical code into general language were identified, which were mostly adapted to the assumed health literacy of the patient profile the simulated patient chose to play, based on an authentic, anonymized discharge summary. As the discharge summaries contained only the gender and the age of the respective patient without any other personal data, the simulated patients had the possibility to personalize their role based on the patient history and the procedures the patient had gone through. The students in the role of the GP benefited from the scenario of knowing the patient for a longer time, which is why it was possible to specifically analyze the deictic elements applied. Four types of personal deixis were detected, most of them not having negative connotations. Some students even addressed the patient by his/her name, treating him/her as a patient who has been treated for a long time. Personal deixis provided language tools which helped the ‘physician’ show his/her empathy to the ‘patient’ as well as his/her ability to guide the conversion in a competent but patient-centered way. Based on the results of the present study, the research question arose whether word sketch differences between medical reports and physician-patient interactions further clarify different connotations of GP’s utterances in the sub-subcorpora. New code-switching strategies might be detected this way, as well as verbal constructions identified serving as a means of introducing diagnosis disclosure.

References

American Academy of Family Physicians (2000). Patient Education. *American Family Physician* 62 (7), 1712–1714.

Balint, M. (1957). *The Doctor, His Patient and the Illness*. London: Churchill Livingstone.

Bigi, S. (2014). Key components of effective collaborative goal setting in the chronic care encounter. *Communication & Medicine*, 11 (2), 103–115.

Bigi, S. (2016). *Communicating (with) Care*. Amsterdam, Berlin, Washington D.C.: IOS Press.

Bourhis, R.Y., Roth, S. & MacQueen, G. (1989). Communication in the hospital setting: A survey of medical and everyday language use amongst patients, nurses and doctors. *Social Science & Medicine* 28 (4), 339–346.

Domonkosi, Á. (2016). Perspective and attitudinal deixis in Hungarian. *Jezyk, Komunikacja, Informacja / Language, Communication, Information* 11, 86–98.

Domonkosi, Á. (2018). The Socio-Cultural Values of Hungarian V Forms of Address. *Eruditio – Educatio* 13

(3), 61–72.
Engel, G.L. (1980). The clinical application of the biopsychosocial model. *American Journal of Psychiatry* 137 (5), 535–544.
Fogarasi, K. (2018). A diagnózis jelentése és jelentősége a beteg szemszögéből [The meaning and significance of the diagnosis from the patient's perspective]. In J. Dombi, J. Farkas & E. Gúti (eds.), *Aszimmetrikus kommunikáció - aszimmetrikus viszonyok.* Bicske: SZAK, 774–804.
Fogarasi, K., Kránicz, R., Halász, R. & Hambuch, A. (2020). Die Rolle medizinischer Wissensvermittlung in Arzt-Patienten-Gesprächen: die Bedeutung des ärztlichen CodeWechsels in Hausärztlichen Konsultationen. *Journal of Languages for Specific Purposes*, 1 (7), 83–96.
Gréciano, G. (2006). Zur Textrelevanz von Phraseologie im Bereich Medizin. In A. Häcki Buhofer (ed.), *Phraseology in motion 1. Akten der Internationalen Tagung zur Phraseologie (Basel, 2004)*. Baltmannsweiler: Schneider Verlag Hohengehren, 219–228.
Halász, R., Kránicz, R. & Hambuch, A. (2021). Die Besonderheiten der Diskurshandlungen zwischen MedizinstudentIn und PatientIn. In T. Schnedermann, Y. Ilg & M. Iakushevich (eds.), *Linguistik und Medizin.* Berlin Boston: De Gruyter, 51–70.
Jobbágyi, G. (2013). Az orvos–beteg jogviszony az új Ptk.-ban. [The doctor–patient legal relationship in the new Civil Code]. *Polgári Jogi Kodifikáció* 7 (3), 15–20.
Kuna, Á. (2016). Personal deixis and self-representation in medical discourse. Usage patterns of first person deictic elements in doctor's communication. *Język, Komunikacja, Informacja / Language, Communication, Information* 11, 99–121.
Levenstein, J.H., McCracken, E.C., McWhinney, I.R., Stewart, M.A. & Brown, J.B. (1986). The patient-centered clinical method. 1. A model for the doctor-patient interaction in family medicine. *Family Practice* 3 (1), 24–30.
Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.
Paterick, T. E., Patel, N., Tajik, A. J. & Chandrasekaran, K. (2017). Improving health outcomes through patient education and partnerships with patients. *Proc (Baylor University Medical Center Proceedings)* 30 (1), 112–113.
Pendleton, D. (1984). *The Consultation: an approach to learning and teaching.* Oxford, New York: Oxford University Press.
Pramann, O. (2017). Einwilligung des Patienten: Rechtliche Details, die Ärzte kennen sollten. *Deutsches Ärzteblatt International* 114 (38).
Safwat, S. (2018). Deixis in the language of nursing. *The IAFOR International Conference on Education – Dubai 2016 Official Conference Proceedings*. Nagoya: The International Academic Forum, 35–46. http://papers.iafor.org/wp-content/uploads/conference-proceedings/IICE/IICEDubai2016_proceedings.pdf
Tátrai Sz. (2011). *Bevezetés a pragmatikába. Funkcionális kognitív megközelítés*. [Introduction to pragmatics. A functional cognitive approach]. Budapest: Tinta Könyvkiadó.
Verschuren, J. (1999). *Understanding pragmatics*. London: Edward Arnold; New York: Oxford University Press.
Wagner, A., Gadebusch-Bondio, M. & Kinnebrock, S. (2020). Informationsflut ersetzt nicht das professionelle Gespräch. *Deutsches Ärzteblatt* 117 (31–32), 1518–1520. https://www.aerzteblatt.de/pdf.asp?id=214882
Weiss, B.D. (2007). *Health literacy and patient safety: Help patients understand*. Chicago: American Medical Association Foundation.
Williams, N. & Ogden J. (2004). The impact of matching the patient's vocabulary: a randomized control trial. *Family Practice* 21 (6), 632–637.
Wood, N. I. (2019). Departing from Doctor-Speak: a Perspective on Code-Switching in the Medical Setting. *Journal of General Internal Medicine* 34 (3), 464–466.
Wright, J.P., Edwards, G. C., Goggins, K., Tiwari, V., Maiga, A., Moses, K., Kripalani, S. & Idrees, K. (2018). Association of Health Literacy With Postoperative Outcomes in Patients Undergoing Major Abdominal Surgery. *JAMA Surgery* 153 (2), 137–142.

The Role of E-health Communities for Older People: A Digital Ethnography

Konstantin Galkin

Sociological Institute, Federal Center of Theoretical and Applied Sociology, Russian Academy of Sciences, St. Petersburg, Russian Federation

E-mail: kgalkin1989@mail.ru

Abstract

In the article, when considering the integration of older people into online health communities, we resort to considering the role of such communities in the lives of older people.

This study contributes to the study of the possibilities and limitations of online communities for older people with chronic diseases living in rural areas. For older people with chronic diseases, such communities can serve as communities for communication, that is, those communities where communication is possible, which is often not enough for older people living in peripheral settlements. The importance of such communities is determined by the fact that it is often difficult for an older person in peripheral settlements due to remoteness to get the necessary advice or recommendation from a doctor, and communication in online health communities and medical forums contributes to obtaining such advice, which is important in the treatment of the disease.

Based on the studied groups of older people, conclusions are drawn and various strategies for using online health communities are typified. In particular, three strategies are identified: the communication strategy, the user strategy and the monitoring strategy. For each of them, the features of the use of online health communities are outlined, as well as the role of such communities in everyday life. Thus, the use and monitoring strategy are characterized by the fact that users — older people — are integrated mainly into online health communities for communication and interaction, while for representatives of the communication strategy, the most significant is the use of online health communities as a communicative platform for discussing various problems existing within rural life and health, and also for therapeutic communication.

Keywords: digitalization, older people, online health communities, rural areas, self-care, communication in online health communities.

1. Background

Contemporary health and healthcare are steeped in neoliberal principles from the hospital and state healthcare systems to the individual patients. One way to integrate neoliberal treatments is e-health illness communities where patients can get the necessary advice on how to treat illness. Individual patients today enjoy autonomy in their healthcare, but this has come at the hidden price of a greater expectation for patients to take personal responsibility for their health and bodies (Hoppania & Vaittinen 2015; Baer 2016).

Online health communities are the form of implementation of diagnostic and treatment, preventive, organizational, and managerial processes in health care through computer technology and active Internet use. Online health communities represent a type of knowledge – not specialized medical knowledge, but the personal knowledge of people who have developed a specific understanding and unique meaning of their illness. Such communication, on the one hand, serves as a means of patient empowerment and, on the other hand, quite naturally undermines the traditional "doctor-patient" hierarchy, as patients become increasingly involved in health self-management and become active consumers of health care and treatment knowledge. Thus, the traditional communicative asymmetry between doctors and patients is breaking down. The role of doctors as experts who interpret symptoms for patients is being blurred by social media, which allows patients themselves to discuss and interpret their symptoms in the public online space. In this regard, some attention should be given to the problem of information imbalance between information production and consumption in online health communities; there is a great need for quality information and descriptions of personal experiences, but only a small proportion of community members tend to

share such information (Conrad, Bandini & Vasquez 2016). Most users turn to online health communities to read what others are writing, and a much smaller number of users actively and regularly engage in discussion (Zhang et al. 2017; Carron-Arthur et al. 2015; Zhao et al. 2022). However, the role of online health communities in the treatment of illness is twofold. On the one hand, online health communities play an important role in the process of self-choice of treatment by the patient, but on the other hand, they can disrupt the treatment process with wrong advice and wrong ways of treating the illness. In this research participation in online health communities is hypothesized to be a buffering mechanism that would facilitate older people with chronic illness coping efforts with illness in everyday life. Individuals coping with extremely stressful life events usually report a need to connect and to talk with people who face the illness and seek out emotional support. Communication in online health communities is suggested to be a comforting aspect of the group support where older people with chronic illness can talk about their illness and talk about the difficulties of experiencing the illness and get the necessary advice. Moreover, participation in online health communities is actually found to be helpful in adaptation to a chronic illness and is conceptualized as a facilitator of patients' ability to integrate the everyday life associated with the experience of chronic illness in individuals' biographies. The aim of my research is to study the interactions and communications of older people in online health communities. Through online ethnography I analysed the features and content of older peoples' posts in online health communities. Additionally, interviews were conducted with the older people that participated in the online health communities. The interview method made it possible to determine the meanings and roles of online communities for the older through their narratives.

2. Theoretical context

It is suggested in medical sociology literature that the interaction between Internet technology and health care consumerism is promoting "information age health care system" in which health care consumers use the information technology to gain access to medical information and control their own health care (Ziebland & Wyke 2012, 219). Accordingly, we can argue that sick role and the patient role has been changing due to changing technological and social trends of societies creating new cohorts of technology using health care consumers (Schnittker 2005).

Symbolic interactionism is a sociological paradigm that can be used to better understand the illness experience of older patients since this paradigm embeds that perception of the self and identity within a social context and links illness associated cognitions to behavioral manifestations. The symbolic interactionist paradigm focuses on individual and group interaction and provides us a framework to understand the link between individuals and their interactions in online communities (Fife 1994).

According to symbolic interactionist paradigm, patients construct their illness in the interaction processes. The information and supportive communication provided by health communities, loved ones and friends play a central role in the formation of these illness representations. Social support from others may offer alternative view of situation and alter the perception of the illness (Fife 1994, 316; Zhao 2005, 387). One of the factors that influence the process of adjustment to an extremely stressful life event, like chronic illness, is to find meaning in the event by the communication with other patients and to incorporate it into personal life course history (Lechner et al. 2003; Bowman et al. 2003, 226; Crossley 2003, 440). Social support can play an important role in finding meaning in the experience by offering alternative views on the chronic illness experiences and by representations of the illness experience (Linley & Joseph 2011, 152; Burrows et al. 2000, 120; Hardey 2001, 31).

Talking about the chronic illness and sharing a supportive context like online health communities is associated with a positive reappraisal of the experience (Beck 2005, 79; Babrow et al. 2005, 31).

Sociological literature classifies social support into three main categories: emotional support, informational support, and instrumental support (Bloom et al. 2001). The first two categories of support are useful to the types of support observed in online health community. In this study I highlighted strategies of communication and participation in online health communities according to these types of social support. Emotional support refers to listening and providing the person with sense of love, respect, sympathy, comforting, encouragement. It also refers to the ability to share one's thoughts, emotions with other members of the community (Schroevers et al. 2003, 378). Informational support refers to provision of information to the members of the community. In the context of chronic illness, recommending a physician to another patient or making suggestions about different medicines or daily regimes based on personal experiences could be examples to this type of support. Instrumental support is support of the members with a concrete problem, e.g. help, doctors advices about hospitals and medical operations or support with medicines, e.g. by sending medicines by post. This is a very concreated type of social support that deals with a concreate problem, not advice as informational support.

Another important concept in the study is the construction of knowledge that occurs in online health communities through the interactions with community members. One major development of this sociological turn in science studies was a shift in empirical focus to non-expert forms of knowledge. I will refer to this eclectic variety of knowledge as *lay knowledge* concept. This knowledge is usually opposed to the traditional laboratory knowledge of scientists which is produced in laboratories and scientists, and doctors have an expert monopoly on such knowledge, according to the work of B. Latour (Latour & Woolgar 1979; Wolpe 2005). For a basic definition of lay knowledge, I use Berger and Luckmann (1966), who famously conceptualized knowledge as the basis for the entire social construction of reality. Parting ways with previous scholars of knowledge who focused only on expert scientific knowledge, they emphasized the importance of "everyday life 'knowing,'" which includes "an assemblage of maxims, morals, proverbial nuggets of wisdom, values and beliefs, myths, and so forth" (Berger & Luckmann 1966, 65). In the realm of health, lay knowledge often combines medical, psychological, and social information to form a colloquial understanding of different health conditions (Krause 2003, 599).

Since the 1980s, in the context of neoliberal changes, «trust in expert knowledge has been replaced by market control. According to this approach, medical care is also a commodity and/or service as any other. Choice in medicine has become increasingly relevant as medicine privatization replaces a public health system with equal access for all. The neoliberal logic of choice is playing an increasing role in both the paid and free segments of medicine and is also actively developing in the context of Internet medicine and Internet use to treat various diseases. Researchers show that descriptions of risk management in the neoliberal model of medicine can use the language of consumerism and neoliberalism, and actual choice practices can be emotionally colored and conform to cultural expectations about trust. As a result, the dichotomous model is questioned by researchers: on the one hand, the passive patient, dependent on the authority of the doctor, unconditionally trusting them and their professional knowledge, which has a touch of sacredness for a profane; on the other hand, the neoliberal informed consumer with a skeptical attitude toward medical knowledge and professionals, making choices consciously and rationally through calculations and "shopping" and participating equally in decision-making (Foley 2008; Glasdam, Oeye & Thrysoee 2015). Thus, from the perspective of the neoliberal model, Internet medicine presents an alternative and choice for older

people.

3. Methodology and empirical base

Online health communities are initiatives that develop from private providers to treat illnesses. The stated purpose of such a service is to provide a platform for discussing various issues regarding the disease and health, as well as problems regarding disease treatment and obtaining alternatives. In the study I employ two methods:

1) narrative interviews with older people;
2) digital ethnography in an online health community focusing on older peoples' postings

The method of narrative interviews is the main one in the study, while the method of digital ethnography is an additional method. I use these methods to consider comprehensively the communication and interaction of older people in online health communities.

This article analyzes 35 biographical interviews with study participants. The study participants are older people living in rural areas. All older people had various chronic diseases. The age range of informants is from 65 to 88 years.

The older individuals who participated in the study had devices that provided access to the Internet. Ten people used personal computers, and 25 people used smartphones. A total of 21 women and 14 men participated in the study. All the respondents lived alone in the villages constantly, but some had relatives who visited them during the summer.

The interviews were analyzed using content analysis. The digital ethnography was an analysis of older peoples' posts in an online health community. It was based on a large online health community that also provides advice, treatment for diseases, and prevention through the e-health system. The posts provided by older residents, residents of villages in Luzhsky District, Leningrad Oblast and villages in the Republic of Karelia were analyzed according to three key themes:

1) everyday life and discussion of everyday problems and treatment of the disease;
2) features of treatment and organization of treatment;
3) problems and features of psychological well-being and features of communication in the community.

A total of 533 posts were selected. Open coding was used for coding the posts and interviews. Content analysis was used to highlight key features and the three themes regarding interactions of rural elderly in online health and living communities and associated limitations.

Online health communities are websites where ill people and doctors convene to collaborate, communicate, and find information about treatment. Communication in online heath communities takes place via forums and symptom logs. Communications on forums consists of "threads" of postings displayed in reverse chronological order. Data source is a large online health community at which several thousand patients meet daily to share knowledge and support. The patients in online health communities have different biographical background and different illnesses. Patients from online health communities are mostly older people from cities and villages of Russia. There are also young people in the communities who are experiencing serious chronic illness. There are also doctors in the communities. Doctors participating in the communities have specialties of general practitioners, cardiologists and surgeons. Basically, the involvement of doctors in the communities is to give advice to patients and moderation of some conversations in the chat. Communities' members register for free, select an anonymous username, and fill out a simple public profile in which they can specify their location, occupation, and interests. Many members report that reading the message boards is a part of their daily routine. Interactions in each message mostly center around the minutiae of a

condition (e.g., symptoms, diagnosis, treatment, remission, etc.), but members also routinely go "off-topic" to discuss, for example, issues related to the communities itself, their personal lives, and the healthcare enterprise as a whole. The communities also feature a chat room, blog, informational resources for various conditions, a bookstore, and a community newsletter. These resources, in addition to the conversations taking place on the message boards, are open to the public.

The three strategies which I highlighted in the study differ in the context of online communities use and interactions on the Internet, as well as in the communication styles of older people. For the representatives of the *communication strategy* the most important is psychological support communication. For the *user strategy* is important using communities to get the necessary information. For the representatives of the *monitoring strategy* the most important are observation in the community and reading of different posts.

4. Findings

4.1. Communication strategy

Representatives of this strategy are usually older people who live alone in rural areas and who have relatively recently started to use the Internet and have a short life experience with the illness. For such informants, the primary importance of online health communities is the ability to communicate, discuss health issues, and participate in communities:

> *Now, if it weren't for the computer, I probably would have been unable to communicate with anyone here at all, I don't really like the neighbors, they are kind of muddy, and here at night and day there is complete silence in the whole village. That's why such online communication helps and rescues a laptop, there you'll at least share it, tell about the illness and everything will become easier. (Informant 1)*

Those older people who chose a *communication strategy* noted the importance of emotional support in online health communities. When writing messages on forums, representatives of this strategy, as a rule, used the emotional style of online communication, with a large number of exclamation points and interjections, preferring to listen to them more quickly than they talked about themselves:

> *To be honest, every time I almost always sit in the online community and almost like this every time I describe my illness enough, the course of my illness, with possible, I don't know, ways out and ways to treat the illness, which I indicate in the descriptions of the illness <...>..(Informant 2)*

That is, when communicating with older people choosing a strategy for presenting the illness, it is important for the community to use often-stoic descriptions of illnesses and possible problems. Not infrequently, I notice gaps related to the fact that in an interview the illness is described much more complicated and the struggle against the illness took on completely different shades than the description of the illness on the forums, that is, corporality in describing the illness was mainly not brought to the online space by representatives of the *communication strategy.* This is associated mainly with offline communication, but the stoic experience of the illness occupied a central plot line when communicating on forums:

> *I can tell you this, how difficult it is, how difficult it is to live with the illness like this every day, I can share with the doctor, but the Internet, in general, is basically not for this from the beginning here either to write a lot of positive or write nothing. I remember the story*

of how I waited for an ambulance for three hours in Luga and almost like that, for three hours and how I didn't die, and wrote like that, they all gave me attention, they gave advice, and I can say so and talked about our medicine, who else to share, and people warned of possible dangers that could happen. (Informant 3)

Confidence in the advice of the representatives of the *communication strategy* is basing on positive advice, which is highly recommendatory, but not edifying. Moreover, since the forum is attended by doctors, representatives of various specialties, representatives of the *communication strategy* say that in this case, the most diplomatic style of advising doctors is important, where the doctors do not demonstrate their expert status:

I don't like when on forums on the Internet, people come to our community, for example, doctors there who think they are demigods, that is, everyone knows and knows how to properly treat the illness. This I do not accept. Probably, the best thing for me is this style I don't know what else is in communication. when I write my posts, and some of the doctors or nurses try to comment on it, but I do it in such a confidential manner without using different ones there scientific words, expressions and so on. (Informant 4)

Also, representatives of the *communication strategy* noted that it is important not only to talk about the experience of the illness itself, but also the stories related to how life happens with a chronic illness in rural areas. Since members of the online community are people from different places and from different regions, it is often discussed in communication such moments as improving life in the village, making infrastructures more convenient, or for example spending hot water, electricity and gas in the village, and the absence of communal infrastructure is very often described by older people in interview narratives as problems caused precisely by the countryside and related to the peculiarities and difficulties in experiencing chronic illness, and then, in addition to discussing chronic illness, an opportunity arose about other problems of life with the illness and get emotional support on completely different issues:

Here we have old people living in the Leningrad region, they will cut down the stove, then the lights will be turned off, something else will happen and then there's nowhere to go if you want to live and live like that, otherwise a bunch of problems just falls on your head, so today I'm telling you how they cut off the light at three in the night and gave it back only at five in the evening, sit like that and don't know how to take a pill (post in online New Health community[27]).

Moreover, as older people with chronic illnesses wrote themselves, sometimes stories about life in the countryside played the role of stoic stories, stories about how difficult it is to just live in a village and live with a chronic illness. This example illustrates the inscriptions used in microsociology when one of the things, for example a game object, ceases to be a game object and acquires new scenarios according to the concept of D. Ball, in this case, the online illness community becomes a community for discussing housing problems, rural ethnography life and descriptions of its difficulties (Ball 1967). Such a strategy is quite typical for representatives of the *communication strategy*, that is, the familiar strategy is broken and the health community is not using for its intended purpose, but as an opportunity to receive emotional support. It should also be noting that representatives of the communication strategy preferred to communicate mainly with people of their age and tried to avoid

[27] The community name is anonymized according to the ethical program of the study.

communication with young community members. That is, in some cases, the communication of the representatives of the communication strategy can be described as quite selective and most often the communication was built precisely with people of their age, with whom it was possible to discuss important and necessary things regarding the treatment of the illness.

4.2. Users strategy

For representatives of the *users strategy*, the characteristic feature is not a periodic violation of the communication scenario, as is typical for a communication strategy, but the following of the general concept of the online health community, or rather a structural branch of this concept, namely the search for necessary and important tips that can be found both in the framework of discussion on forums, and in the framework of communication with other people, something for which the community itself was originally conceived. As a rule, unlike representatives of the *communication strategy*, representatives of the *users strategy* use to go to online health communities less often, and went in when it is necessary to solve some health problem or in cases of sharp remissions or a sharp deterioration:

> *I don't go there very often, but sometimes I go there just to find the advice I need, I recently found the necessary prescription and started using this medicine and if it weren't for the community, where could I find it, but here it's easy just found. (Informant 6)*

Representatives of the *users strategy* preferred to communicate on the forums only when it was really necessary to provide others with information or to ask for advice. As a rule, representatives of the *users strategy* were those older people who live with family or relatives and one of the explanations for the quick use of online health communities was that they just briefly focused on information because they had family and relatives for general communicative purposes. It is also important that the representatives of the *users strategy*, as a rule, communicated with neighbors and maintained fairly close relations with them, which, of course, had an impact on how communication took place in communities and on understanding the instrumental role of online communities:

> *Two years ago, when my daughter was still living with us, she bought me a laptop that was not expensive, HP, and since then, it happens, no, no, and I'll look on the Internet, but I entered the community when I was looking for my record, well, there's a long series of links to the doctor, and as a result, a link to the person I need and then I read, liked the material, liked that the therapist speaks there, generally writes about illnesses and treatments, well, in general, as a result, I joined the community. (Informant 7)*

Representatives of the *users strategy* noted that they were guided by the Internet community, as a community of experts, and trusted the opinion expressed there by both doctors and other patients with similar illnesses. Often representatives of this strategy say that there was much more useful content for treating the illness in the online community than simply in directories or on forums on the Internet. Therefore, communication in online health communities was important in terms of finding new opportunities and ways to treat the illness. In consequence, psychological help could move to a space outside the online health communities and the transit was accomplished just on the basis of trust in the e-health community. For members who use this strategy it is important not just to read the tips but also to implement them, follow these tips.

For representatives of the *users strategy*, it is typical to follow the strategy prescribed by the e-health community. To get the necessary advice or recommendation on how to treat the illness is more

important than to communicate with other people. That is, long communication and long stories about the illness can make older people leave the online health community and withdraw from the chat itself, and usually then they do not enter other communities to communicate and discuss the illness and their condition in them. One of the important research observations was that for older people who chose a *users strategy* satisfaction with rural life was characteristic, in contrast to representatives of the *communication strategy* and many of its aspects. Usually, representatives of the *users strategy* asked for advice on medicines, replacing medicines with generics, and the possibility of using certain medicines and their benefits.

4.3. Monitoring strategy

For representatives of such strategy, it was important to observe how communication occurs in forums, in online health communities. At the same time, representatives of such a strategy did not participate in discussions, but simply read forums of online health communities without being a member of this group. This use of online health communities is argued by representatives of the *monitoring strategy* as reluctance to engage in dialogue with other people. Many of the representatives of the *monitoring strategy* preferred to remain silent and not talk about their illness, considered it unnecessary, and therefore tried not to participate in discussions on online forums:

> *I'm just there, I'm reading various communities, that is, I'm just reading without any comments, I don't know any comments or participation, but sometimes it's useful and sometimes you can find a lot of good and necessary information there, but I'll say – so late . (Informant 9)*

That is, it is often important for representatives of the strategy to read and receive the necessary information, but unlike representatives of the *users strategy*, these informants did not consider it necessary or important to ask questions. The online health community is something of an "encyclopedia", as one of the informants once called it. That is, it was important to be able to read something and bring out something for yourself.

Representatives of the *monitoring strategy* have different backgrounds and different attitudes to life in rural areas, both positive and negative, and the spectrum of this ratio could vary significantly.

5. Conclusion and discussion

Rural patients, restricted solely to public medicine, act as citizen-consumers. They choose the service that suits them by reading messages within online health communities and communicating within online health communities. Thus, the logic of choice that is typical of the neoliberal approach to medicine allows the use of online health communities by older people to obtain alternative advice and options in chronic disease treatment and the strategies we have highlighted.

The strategies of *users* and *communication* considered in this study show us that the online health community itself (in addition to the practical benefits of communication and the ability to speak emotionally, talk about your illness, or simply read the necessary information and ask for information about the necessary prescription) creates opportunities for the manifestation of patient agility and the opportunity to participate in a dialogue and discuss various problems and issues.

The results showed that older people trust in knowledge and so-called unproven services. They searched for advice on the basis of which it was possible to provide or not to provide necessary services in the future (credible goods) in online illness communities, which are created in such online

health communities, if doctors participate in discussions on forums This indicates that the fluidity of such communities can be regarded both positively and negatively by users. Therefore, I would not like to take the position of a judge who assesses the possibilities of online health communities for older people with chronic illness who live in rural areas. However, I will insist and remain in the position of a researcher who is able to present the variability of various strategies for using this fluid technology – online health communities –and understand those ways and rules on how to use the discussions and interactions, as well as the potential of communities.

However, one of the important conclusions, in my opinion, at this stage of the study, is how older people translate rural life and the difficulties of rural life into online health community spaces; this is largely determining by two strategies: *communication strategy* and *users strategy*. For representatives of the *communication strategy*, the lack of communication and interactions in the rural community is transferring to the space of the village and, to some extent; the illness communities replaces the village community exactly the opposite. That is, the village community, with the strong social connections and advice that is present to it, is changing, the collectiveness is disappearing and not the advice is important, but rather the emotional and therapeutic support that e-health community provides.

Representatives of the *users strategy* use online health communities rather as a tool in the treatment process. This tool is important and necessary based on the general dissatisfaction with the doctors and the inability in the countryside, even when talking with neighbors or relatives, to ask for the necessary advice in the treatment of the illness. At the same time, lonely or not lonely living is an important factor in using online health communities, which is also important for what role such communities will play in life with chronic illnesses. That is, the current technologies of online health communities represent many opportunities and are expressing in the mass of views on how you can manage your illness. Often for residents of rural areas, such management and the choice of the necessary method is perhaps the only opportunity to consult with someone or ask for the necessary advice on taking medications or changing the schedule of taking medications in everyday life.

Another important aspect that deserves attention in the continuation of the future discussion is the features of corporality and the agency of their manifestations during communication and interactions in online communities.

6. Acknowledgements

The research was carried out at the expense of the grant of the Russian Science Foundation No. 22-18-00461 (https://rscf.ru/project/22-18-00461/).

References

Babrow, A. S., Kline, K. N., Rawlins, W. K. (2005). Narrating problems and problematizing narratives: Linking problematic integration and narrative theory in telling stories about our health. *Narratives, Health, and heeding: Communication Theory Research, and Practice*. New Jersey: Lawrence Erlbaum Associates, 31–60.

Baer, H. (2016). Redoing feminism: Digital activism, body politics, and neoliberalism. *Feminist media studies*, 16 (1), 17–34.

Ball, D. W. (1967). Toward a sociology of toys: Inanimate objects, socialization, and the demography of the doll world. *The Sociological Quarterly* 8(4), 447–458.

Beck, C. S. (2005). Becoming the story: Narratives as collaborative, social enactments of individual, relational, and public identities. *Narratives, Health, caul healing: Communication Theory, Research, and Practice.* New Jersey: Lawrence Erlbaum Associates, 61–81.

Berger, P. & Luckmann, T. (1966). *The Social Construction of Reality: A Treatise in the Sociology of*

Knowledge. New York: Doubleday, 249.
Bloom, J. R., Stewart, S. L., Johnston, M., Banks, P. & Fobair, P. (2001). Sources of support and the physical and mental well-being of young women with breast cancer. *Social science and medicine* 53 (11), 1513–1524.
Bowman, K. F., Deimling, G.T., Smerglia, V., Sage, P. & Kahana, B. (2003). Appraisal of the cancer experience by older long-term survivors. *Psycho-Oncology* 12, 226–238.
Burrows, R., Nettleton, S., Pleace, N., Loader, B. & Muncer, S. (2000). Virtual community care? Social policy and the emergence of computer mediated social support. *Information, Communication & Society* 3(1), 95–121.
Carron-Arthur, B., Ali, K., Cunningham, J. A. & Griffiths, K. M. (2015). From help-seekers to influential users: a systematic review of participation styles in online health communities. *Journal of medical Internet research* 17 (12), e4705.
Conrad, P., Bandini, J., & Vasquez, A. (2016). Illness and the Internet: From private to public experience. *Health* 20(1), 22–32.
Crossley, M. L. (2003). Let me explain: Narrative emplotment and one patient'sexperience of oral cancer. *Social Science and Medicine* 56 (3), 439–448.
Fife, B. L. (1994). The conceptualization of meaning in illness. *Social Science & Medicine* 38(2), 309–316.
Foley, E. E. (2008). Neoliberal reform and health dilemmas: social hierarchy and therapeutic decision making in Senegal. *Medical Anthropology Quarterly*, 22(3), 257–273.
Glasdam, S., Oeye, C. & Thrysoee, L. (2015). Patients' participation in decision-making in the medical field – 'projectification'of patients in a neoliberal framed healthcare system. *Nursing Philosophy*, 16 (4), 226–238.
Hardey, M. (2001). 'E-health': the internet and the transformation of patients into consumers and producers of health knowledge. *Information, Communication and Society* 4 (3), 388–405.
Hoppania, H. K. & Vaittinen, T. (2015). A household full of bodies: Neoliberalism, care and "the political". *Global Society* 29 (1), 70–88.
Krause, M. (2003). The transformation of social representations of chronic disease in a self help group. *Journal of Health Psychology* 8 (5), 599–615.
Latour, B. & Woolgar, S. (2013). *Laboratory life: The construction of scientific facts*. Princeton University Press, 293.
Lechner, S. C., Zakowski, S. G., Antoni, M. H., Greenhawt, M., Block, K. & Block, P. (2003). Do sociodemographic and disease-related variables influence benefit finding in Cancer patients. *Psycho-Oncology* 12, 491–499.
Linley, P. A. & Joseph, S. (2011). Meaning in life and posttraumatic growth. *Journal of Loss and Trauma* 16 (2), 150–159.
Schnittker, J. (2005). Chronic Illness and depressive symptoms in late life. *Social Science and Medicine* 60, 13–23.
Schroevers, M. J., Ranchor, A. V. & Sanderman, R. (2003). The role of social support and self-esteem in the presence and course of depressive symptoms: a comparison of cancer patients and individuals from the general population. *Social Science & Medicine* 57(2), 375–385.
Wolpe, P. (1985). The maintenance of professional authority: acupuncture and the American physician. *Social Problems* 32(5), 409–424.
Zhang, X., Liu, S., Deng, Z. & Chen, X. (2017). Knowledge sharing motivations in online health communities: A comparative study of health professionals and normal users. *Computers in Human Behavior 75*, 797–810.
Zhao, S. (2005). The digital self: Through the looking glass of telecopresent others. *Symbolic Interaction* 28 (3), 387–405.
Zhao, Y., Chen, K., Peng, J., Wang, J. & Song, N. (2022). Diverse needs and cooperative deeds: Comprehending users' identities in online health communities. *Information Processing & Management*, 59 (5), 103060.
Ziebland, S. & Wyke, S. (2012). Health and illness in a connected world: how mightsharing experiences on the internet affect people's health? *Milbank Quarterly 90*(2), 219–249.

Appendix: Informants

Informant 1: male, age 68, single, third degree of disability, heart attack.

Informant 2: female, age 78, family, third degree of disability, vessels illness.

Informant 3: female, age 65, single, second degree of disability, stomach cancer.

Informant 4: female, age 85, single, third degree of disability, heart attack.

Informant 6: female, age 71, family, second degree of disability, blood-stroke.
Informant 7: male, age 80, single, third degree of disability, spinal column illness.
Informant 9: male, age 67, single, second degree of disability, carcinoma of lung.

Reflections on Digital Ethnography and Digital Realms of Young People

Kristiina Korjonen-Kuusipuro & Sari Tuuva-Hongisto & Päivi Berg

South-Eastern Finland University of Applied Sciences

Email: kristiina.korjonen-kuusipuro@xamk.fi

Abstract

In this article, we discuss digital ethnography and reflect how it allows researchers to better understand the digital realms of young people and to study their digital everyday lives. We draw on quantitative and qualitative research material produced in 2021–2022 on the digital everyday life of young people in Finland born in 2005 and 2006. The COVID-pandemic changed our data collection, which was moved to online environments. Central to our research was defining the field, building the interaction and rapport between researchers and participants, and implementing observations in online environments, because young people use a variety of platforms. Young people were also strict about their privacy, and thus continuous ethical reflection is an important part of our digital ethnography.

Keywords: digital ethnography, digital landscapes, youth, Finland

1. Introduction

Researchers (e.g., Bennett & Maton 2010) no longer agree with the previously commonly held view of young people as digital natives (Prensky 2001a; 2001b). By suggesting that young people form a homogenous group where all are automatically competent in digital society, we ignore the fact that they are also confronted with changing circumstances and transforming opportunities. Furthermore, young people's online activities and interests are diverse and individual, and their digital skills and motivation vary widely (e.g., Eriksson & Tuuva-Hongisto 2019). We have identified a gap in understanding this diversity. Filling this gap requires research approaches and methodological choices that study young people's individual experiences, everyday practices and lifestyles and consider the lives of young people online and offline as entanglements, not as separate entities.

In this article, we examine the conduct of digital ethnography during the pandemic and reflect on how digital ethnography allows researchers to better understand the digital realms of young people and to study people in their digital everyday lives. The article is part of Dequal research project in which we examine the inequalities of digitalization on the lives of young people in Finland.

Our article draws on quantitative and qualitative research material produced in 2021 on the digital everyday life of young people born in 2005 and 2006. The quantitative material offers background information and consists of an online survey (n=406), which was distributed through schools in three different areas in Finland. The qualitative material consists of individual interviews with 28 young people, two group interviews, and an online observation of the people followed by the participants on social media, as well as the Instagram behavior of some participants. In this methodological article, we use our data collection use our data collection as an example and illustrative case of digital ethnography.

We begin our article with an overview of online ethnography and its various definitions. Next, we describe our own research during the Covid pandemic among youth and reflect on the specific features of digital ethnography. We conclude by illustrating shortly the benefits and challenges of digital ethnography in our study.

2. Messy Web, Multidimensional Ethnography

As a qualitative, exploratory approach, ethnography is more than just a research method. Ethnography

belongs to the realm of interpretive science, it is based on first-hand (participant) observation of (online/offline) social practice, and it has a special focus on everyday lives, choices, and their meanings (Beuving 2020). Central is the so-called 'thick description' and writing about people (Geertz 1973), as well as conceptualizing and theorizing cultural phenomena based on fieldwork, and reflecting on the research process and knowledge production. The role and interaction of the researcher with the participants is essential. As a multimodal approach, ethnography can include a variety of methods, analysis, and interpretation, and is not just a description of the people or community being researched, but an important part of understanding the field and its contexts and connecting the field to broader historical processes. (Geertz 1973; Hine 2015; Koskinen-Koivisto, Lähdesmäki & Čeginskas 2020).

There are several ethnographic trends that have emerged for the study of online communities; cultures that have moved online or developed there. The research literature discusses, for example, virtual ethnography (Hine 2000) or virtual anthropology (Weber 2015), netnography (Kozinets 2009), media ethnography (e.g., Postill 2009; Horst, Hjorth & Tacchi 2012), digital ethnography (Pink et al. 2016), social media ethnography (Postill & Pink 2012) and networked anthropology (Collins & Durington 2014). Although the trends are close to each other, they define their relationship to technologies, research topics, researcher position and research practices in different ways.

The diverse designations of online ethnography have a connection to the history of the Internet and the study of different phases of this history. Pioneering research examined cyber cultures, early Internet content, and the mobile cultures brought about by new technologies, computers, the Internet, and cell phones. Online communities and identities were also examined, and for example, sociologist Sherry Turkle (1995) analysed in her seminal work how "life on screen" affected identity, community, and gender. During the 21st century, the emphasis shifted to the ubiquitous impact of technology. For example, netnographic approach and process, developed by Robert Kozinets (2002; 2009), involved reviewing potential online cultures, gathering research material through more or less participatory observation, making reliable interpretations after analysis and long-term observation, ethical evaluation, and providing opportunities for research participants to comment. (Kozinets 2009; Rokka 2010.)

Looking back, it is easy to see how the developments in technology have shaped researchers' interests. From the mid-2000s onwards, online research focused on Web 2.0 thinking, the backbone of which was formed by social media and various applications such as Facebook (Miller 2011) or social games such as Second Life (Boellstorff 2008) and their cultural reviews (Caliandro 2018). In 2010s phenomena such as datafication, big data, and algorithms became the subject of research (see, e.g., van Dijck 2014; Lehtiniemi and Ruckenstein 2019; Lugosi & Quinton 2018). Researchers have also discussed widely issues of inequalities, power relations, artificial intelligence, and ethical aspects of technology in the online world (e.g., Hine 2015; Richardson 2015; Helsper 2021).

Today, ethnography conducted in online environments is defined as a "multimodal qualitative research method that seeks to understand digitally mediated interactions, communities, identities, and norms by looking at the interfaces of physical and virtual spaces, the interactions between them, the materiality, texts, and bodies that function in these spaces" (Standlee 2017, 329). The key difference is whether digitality is viewed from the real world as an object or an artifact or whether it is studied as a culture, thus blurring the distinction between the real world and the virtual (Hine 2000; Caliandro 2018). Exploring people and understanding digital spaces adds value to the study of physical communities and interactions (Hallet & Barber 2014.) Practices, ways of doing things that are taken for granted, are often too mundane and difficult to describe or even see (Hine 2015, 8.). Ethnographic

research helps to understand the fragmented worlds of digital cultures, communities, and the digital everyday life.

In our study, we define digital ethnography as a way to understand the meanings of digital technologies and the social and cultural practices of digitality in Finnish society and people's everyday lives. Further, we understand digital ethnography as a means to utilize digital tools at different stages of research. (Pink et al. 2016; Standlee 2017; Abidin & de Seta 2020). Previous research has shown a difference between online ethnography and mobile ethnography, i.e., the digital world that opens up via mobile phones and the use of social media (e.g., Ito, Matsuda & Okabe 2006; Thulin & Vilhelmson 2007; Ito et al. 2009; Hine 2015). Digital ethnography combines these perspectives, as the digital world is seamlessly intertwined into a whole comprised of computers, the Internet and smartphones, information and communication technologies, and their various applications, uses, and practices.

Due to the inconsistency of the online worlds, the most important guideline in our research is reflexivity, a continual process of self-reflection and ethical reflection on how we best obtain information about the phenomenon we study, how we understand the diversity of this phenomenon, and how we can present this diverse information as appropriately as possible (Davies 2002; Licheterman 2017; Markham 2021). To leave space for research-enriching improvisation and surprise (Cerwonka & Malkki 2007), we have tried to avoid overly categorical divisions and approached the web as a specific space for young people to be with their peers without strict plans and schedules. In this kind of space, young people develop knowing which geographer Noora Pyyry (2019) and her colleagues define as “hanging out knowing”; thinking that arises in lingering encounters that produces an understanding of a person’s place in the world (El Founti, Leino & Pyyry 2021).

3. Digital Ethnographic Research during the Covid Pandemic among Youth

3.1. Construction of Field and Interaction in Digital Environments

Ethnographic research usually begins by locating the field, defining it, and negotiating access to the field. The field is rarely a permanent space, but it is constructed situationally, temporarily, often physically located in different places, scattered. In digital ethnography, the field is often formed as a combination of physical and online spaces, and research is characterized by a constant redefinition of the field. (Burrell 2009; Hine 2015; Standlee 2017). Digital field sites can include anything that the web is made of, for example platform infrastructures, images, videos, texts, social relations, user behavior and so on. (Góralska 2020, 47).

Our research focuses on ninth graders (born mainly in 2005/2006) in three different regions in Finland. The production of the research material started with a survey on the use, inclusion, and place of residence of young people. We targeted youth living in both urban and rural areas in Finland, because we presumed that home regions are part of concrete opportunity structures for young people. The ninth graders were chosen as a target group, because they are at the start of their transition years for their life-courses concerning education and growing up, approaching emerging adulthood in economically and socio-culturally different environments. In addition, in Finland, ninth grade is the last year of comprehensive school for all students. After that, the education paths of young people separate.

The survey was planned to be conducted in collaboration with schools, but due to COVID-19 the collaboration proved more challenging than planned. Some schools switched to distance learning during the first implementation of our survey in spring 2021. From contacts to schools, it soon became

evident how both the teachers and the pupils were stressed due to the exceptional circumstances and coping with the pandemic. School staff were also tired of all research requests, and the additional work did not inspire them. Although we offered to hold questionnaire-related sessions for pupils ourselves, not all schools responded to our messages, and some refused to participate in the research. Access to the field was challenging and difficult. After several non-responses, we modified our plans and received 406 (53.7% female, 41.3% male, 2.2% non-binary) responses from 25 municipalities to our survey with the contact information of some volunteer interviewees. We also marketed the survey through social media by purchasing visibility on Instagram. This spawned a few respondents and one interviewee.

Although the subject of the research project is the digital everyday life of young people, our purpose was to meet young people face-to-face, chat with them, interview them and observe their use of digital media and devices. During our research, we faced a new challenge, as the COVID-19 pandemic changed our data collection plan almost completely. The only option was to transfer the entire study online. The flexibility of the ethnographic research method and the creativity of the researchers had to be fully implemented within a short timeframe. Relocating the data collection to an online environment also raised questions about digital ethnography as being conceptually diverse, sometimes even confusing, and fuzzy. Ethical aspects also raised timely questions and required careful consideration.

Traditionally, ethnographic approach involves close interaction between the researcher and the participants and establishing rapport between the researcher and the participants is of vital importance (Atkinson et al. 2001). As the development of confidential relationships does not take place in an instant, time should be used at the data collection phase for researchers and participants to get to know each other. In our study, we considered that every message we sent to young participants built interaction, and therefore we also attempted to respond to their questions and comments very quickly, even though it may have been a Sunday night. We gave as much decision-making power as possible to young people themselves, and emphasized at every stage that participation is voluntary, and the ways of participation can be modified according to their wishes. The first contact was by email, to which young people did not often respond. If we had the participant's phone number, the next message was sent by WhatsApp.

During autumn 202–winter 2022, we interviewed 28 young people and conducted two group interviews. The interviews were conducted with Teams or WhatsApp, and recorded. The young people themselves were able to choose the application they preferred. What was different to vis-à-vis interview was the lack of face-to-face interaction. We tried to compensate this lack by using a computer camera, but the participants were free to choose whether to keep the camera on during the interview. For us researchers, the interviews were clearly more meaningful when the camera was on, because this made it possible to see the young persons' facial expressions, and their embodied reactions to the questions and conversation. Sometimes technical reasons, such as a poor network connection, prevented the use of a camera. At the end, we felt that online interviews were sometimes even easier for the participants when they did not have to meet the researcher face-to-face (see also Hine 2015, 164–165). Our interviews concentrated on three themes: agency, online practices, and place. We discussed for example what kinds of online activities young people have and what kinds of daily rhythms they have. We also discussed how belonging to groups and friendships are formed online and what kind of pressures acting online has created. We asked our participants to choose and show us some of their online posts and we also asked who they follow online. We also wanted to know what they would change if they had the power to modify online worlds.

3.2. Ethical Considerations

When entering the field, we soon realised that we needed to re-consider the key principles of good scientific practice. Even though the principles of reliability, honesty, respect, and accountability (ALLEA 2017) and the guideline that research should not harm subjects in any way (Murphy & Dingwall 2001) are clear, their implementation in everyday research requires constant attention, context-based interpretation of social situations and human sensitivity (Korjonen-Kuusipuro & Kuusisto-Arponen 2019). In digital ethnography, ethical aspects seemed to be even more blurred and tricky, than in traditional ethnography (Góralska 2020; Thompson et al. 2021).

In our own study, the ethical "hot spots" of digital ethnography were related to applying for ethical vetting, entering the field, negotiating participation, obtaining consent, and being aware of and possibly dismantling different power relations. It was particularly important to consider the publicity of digital spaces. This was directly related to the ethics of following young people's social media channels and negotiating with young people about this (see also Standlee 2017).

The consent for the research and the privacy statement we sent to all participants beforehand by e-mail. In the interviews, it became clear that not all had read these before the interview. Therefore, the consent and the privacy statement were reviewed orally at the beginning of the interviews, and the young people were asked again if they wanted to participate in the study. We also told our participants that even though we can use direct quotations, individuals cannot be identified, because we will use pseudonyms and placenames and other possible identifiers will also be removed. After the end of the project both the qualitative and the quantitative digital research data will be fully anonymized. If possible, the data will be archived so that it will be usable in forthcoming research projects and thesis.

The ethics of digital ethnography was also associated with the visibility of the researcher's presence in digital environments. As with ethnographic research in general, a researcher doing digital ethnography can choose whether to place more emphasis on participation or observation in conducting participatory observation. Researchers may be active commentators and participants on various online platforms, or they may be observers in the role of more passive observers who do not make themselves visible to the person being observed. Invisible observers, the so-called "lurkers", have been considered ethically questionable in some studies, and in our own study, for example, such ethically questionable researchers could have followed young people's Instagram accounts without telling them about it, but this would have been ethically unacceptable. On the other hand, researchers can be followers, for example in the open YouTube video platform, to observe in the same way as researchers who observe the hanging out of young people in general, and do not influence the field they are researching (see also Standlee 2017).

To study young people using digital ethnography opens significant possibilities, but it requires accuracy from researchers. The social media channels favored by young people contain a lot of content produced by the young people themselves but utilizing these requires careful and detailed ethical reflection from the researcher already at the beginning of the research. The researcher must tell young people very clearly how the observation will be carried out and how the research material based on the findings will be used, when the observation begins and when it ends. For example, we sent messages to those young people we followed in Instagram that our observation period has ended. If the situation changed, young people needed to be reinformed. Therefore, having and maintaining continuous contact with the young people was very important, if the young people themselves are willing to do so.

3.3. Defining Online Privacy

Young people scattered online are difficult to reach and can be very precise about their own private space (differences between private and public, for example see boyd 2007). This was also the case in our study. We would like to emphasize that researchers must be aware of their power and respect young person's decision, even if the desire to hang out with the young people online is strong and easy to implement in the online environment. Time must be allowed for the trust between the researcher and the research participant to be built. As trust grows, the young person may later open up about their own activities to the researcher or find alternative ways to open their own use of the network without feeling relinquished. Young people use a variety of platforms to achieve their own "hangout goals" and privacy settings vary depending on the uses of the platform. However, young people participating in our study seemed to be aware of where and how they operate in their own online environments (see also Pitt et al. 2021). Additionally, this role was reflected in the fact that young people could exclude researchers from their use of the network: "*Uh, wait… Let me see what would make sense. You can't follow my TikTok, I don't want to let you in there. [laughs].*" (Maria) For Maria, TikTok clearly appeared as her own, adult-free space.

During our research, we soon realised that our young participants were very careful about their own privacy (see also boyd 2007). Most often their use of digital devices was to communicate with their own friends by using the Snapchat, for example. The other social media accounts (such as Instagram) were private instead of public. They commented on their own roles in the following way:

> *I don't feel I need to introduce my own face in any social media platform or such. (Kristian)*
>
> *I am not such a public person, […] but I have been wondering should I download Instagram, but I am still kind of thinking about it. (Sofia)*
>
> *I don't want that people at my school or even other people in Finland can find me. Therefore I have not advertised it any way. Because I don't want any such person, who can identify me, to come and say "Hey you suck as a person" or something else, because Finns have a very narrowminded attitude towards these things – these are not so big issues here yet. (Lotta)*

Sofia's quote describes the deliberation of young people about what apps they do and do not use. The last quote is from a young content creator who wanted to completely hide her personality; she did not use her own name, she had told no one where she was from, and not even her siblings or friends knew what she was doing.

Some of our participants allowed us to follow their private account on Instagram, after which we researchers considered how we could write about our findings without people being recognized or without harming participants in any other way. In this context, the power relations between the researcher and the participants also came to the fore. How, as researchers, we justified our desire to follow them on social media in general, while assuring them that they did not have to consent unless it felt good. This ethical reflection required us researchers to have social sensitivity, openness, the ability to listen and the tendency to genuinely care from the young person's point of view (Thompson et al. 2021). Indeed, only individuals who clearly expressed an interest in participating in this part of our study were selected.

An ethnographer conducts research with the entire body while experiencing and recording multisensory experiences in the field (e.g., Koskinen-Koivisto & Lehtovaara 2020). In our digital

ethnography, using the camera in the interview situation created a private space. The young person's own home and room was a safe place to meet a researcher. At the same time, camera also made visible the young person's way of being in the digital space, on the one hand open, and on the other hand, hidden from the gaze of the camera and the dialogue. Interview situations online highlighted the specificity of the digital encounter. In the shelter of facelessness, many young people were very open, and this could even encourage them to talk about even the most sensitive and difficult things (e.g., bullying or sexual harassment) that would not have been possible to talk face-to-face (also Gibson 2022). Thus, a participant can feel heard and understood online as well as in a face-to-face encounter. We must also stress that even though a researcher is not (usually) a therapist, we must be ready to act as responsible adults when a young participant talks about sensitive and difficult issues. Every young person has a right to be heard and seen, and thus researchers need to show empathy, and know where a young person can get support and how to guide them forward (for more about online interviews and sensitive topics, see Gibson 2022).

3.4. Observation Online and Entangled Reality of Digital Ethnography

The different dimensions of digital ethnography, its opportunities and challenges, are strongly intertwined, and in the constant state of becoming. The web is fuzzy for both its researchers and users, and is very often described as fragmented, fluid, and messy. Fluidity involves rapid change in both different social and technological environments, and technological capability of the researcher is linked to these changes. Doing digital ethnography is not only lying on the sofa and hanging out in Facebook, YouTube, Twitch or other platforms (see also Góralska 2020).

Our online observation was built on the experiences of young people and our aim is to trace the digital landscape of youth (about digital landscapes see Webster, Svalastog & Allgaier 2020). In practice, this has meant doing ethnographic observations on YouTube channels our participants presented to us. We traced young people's digital everyday lives and interactions through assignments attached to interviews. Before the interview, we sent the young people assignments where they had to choose three of their own "postings" made online and present them to the researchers. They also had to introduce the three channels/content creators they watched online. We also asked young people who are more active in using Instagram to follow their account through the [anonymized] Instagram account of the project. We wrote a field work diary about Instagram tracking. This resulted in a field work diary that contains written notes, screenshots, links to other media content, such as news stories, media channels on other platforms, or even Wikipedia articles, which contained information on the content creator(s). Additionally, the description from the interview was combined with the field work diary.

The online observation concretizes the hanging-out-knowing, which stems from the idea of walking together. Geographers Elena El Founti, Ilari Leino and Noora Pyyry (2021, 140) write that "the recognition of hanging-out-knowledge requires the abandonment of the illusion of linearity and systematicity, because the knowing human subject is seen to be formed in the event of knowing". Noora Pyyry (2019) defines hanging-out-knowing as "an ongoing process that takes place in everyday encounters through negotiations responsive to landscapes, of which one is a part". We understand that these landscapes can also be virtual. Due to the pandemic, we could not hang out with young people face-to-face, so by hanging out through the media channels presented to us by our young participants, it became possible for us to explore their experiences, and in a way, virtually walk alongside them and follow the "footprints and hints" young people gave us within the framework of the interview situation (cf. Robards & Lincoln 2017).

What kind of footprints and hints did the young people give? According to our research, the virtual landscape where young people hang out is very fragmented and differentiated. Almost all the young people we interviewed, followed different content creators and producers, and were interested in different kinds of content, although they always mentioned it as being very popular and well-known. As we gave the task to our participants beforehand, they had carefully selected what they want to show us. This selection can tell us about identity building, but also building ethnicity and race (e.g., Nakamura 2008). Most of the young people also watched several different channels, as can be seen from the following interview excerpt.

> *So Trippydraws makes everyday life videos and updates her everyday life. She draws, let's say space - themed pictures, among other things. (...) I have followed her more on TikTok, so I am not so familiar her Youtube to the content. She has also made some songs. I have listened some of them. [...] Here are a couple of videos where she paints. I think she's just so good at what she does. (Maria)*

Contrary to our preunderstanding, young people themselves produced very little content online. In addition to private communication, young people's digital practices seemed to focus on following the "content" of different parties, mostly friends and acquaintances and more widely known content producers. Our ethnographic observation focused on more widely known content producers that young people told us they were following. We did not seek to analyze the content of these but explored them as if they were a window into the digital lives of young people. We aimed to share their views on the digital practices that are important to them and explore the digital landscape (Webster, Svalastog & Allgaier 2020) of young people. Here we also had to challenge ourselves and reflect on our own perceptions. We experienced that we were indeed approaching a foreign culture, getting to know and understand something new and unfamiliar to us, as is characteristic of classical ethnographic research. It was also a matter of gaining access to experiential and embodied knowledge, in which case it is important to accept all kinds of narratives and experiences as part of the research (Korjonen-Kuusipuro & Kuusisto-Arponen 2019).

> *Well, on Instagram, maybe like that, there are videos of Finnish football technique coach (...), which I've been watching myself. And then on Youtube then, and on Instagram too, I like watching about bodybuilding, and sports in general. Actually, this is all I watch. (Matias)*

When we started to hang out on channels favoured by our participants, a strange world, a new culture, opened for us. In the spirit of posthumanism, we emphasize that we no longer talk about the researcher's gaze at a particular limited area but understand knowledge as something constantly in-the-making (e.g., Pyyry 2019).

Many methodological descriptions of digital ethnography still emphasize the interfaces of physical (offline) and virtual (online) environments and make a distinction between the real world and the virtual world. For this reason, it would be useful to consider the significance of these emphases. Is the distinction between different worlds perhaps pointless? Tom Boellstorff (2008) stated that technology should not be understood as separate from man or humanity. Although he himself refers to the "real world" (physical world or actual world) and the "virtual world" (online world). However, he does not contrast these worlds, but sees them as complementary, focusing on the examination of different aspects of humanity in these different worlds. (Boellstorff 2008, xxvi, 248–249.) An online ethnographer does not need to distinguish between the virtual and real worlds but is largely about operating in many different social environments (boyd 2007). It is possible to see the

real and the virtual as a continuum of each other, simultaneous and interdependent (Sumiala & Tikka 2020) or as entanglements, because as one of our interviews shows:

> *[…] Well [checking his phone] it says here that I am 5 hours 35 minutes online daily. But most of this time is used to watch a film, while doing something else (electricity stuff). I like to watch something at the same time. (Julius)*

4. Concluding Remarks

In this article, we defined digital ethnography as a useful way to understand the social and cultural practices of online communities and the diverse meaning-making processes and meanings of the network in the daily lives of young people. On the other hand, digital ethnography is a way for us to take advantage of digital tools at different stages of research. In our research, we scrutinized the characteristics of digital ethnography and examined how it can best highlight the diversity of the digital everyday life of young people today. Although digital technologies are ubiquitous; the use of devices and applications permeate a young person's everyday life, the practices, customs, and meanings associated with it are highly fragmented and diverse.

The global pandemic situation forced us to do more technology-driven research than we were originally prepared for. For this reason, the interaction between researchers and the young people participating in the study was built on WhatsApp and e-mails. Through interviews with Teams and WhatsApp, we found that online encounters could be even easier for young people than face-to-face encounters. However, even though the young people were very open in the interviews, they were also very careful about their own privacy, and before the interview we often communicated about, for example, the need to record the interviews and keep the cameras on. Researchers making observations were not allowed everywhere and young people thought very carefully about what they wanted to show to the researchers. In our research, for example TikTok was considered their own space to which the researchers were not welcome or permitted to observe.

In our research, we also used the concept of hanging out by aiming for lingering encounters with young people and hanging out on their favourite digital platforms. Even though, the world of research today demands often fast research results which does not actually support these kinds of lingering encounters, these seemed to increase our understanding of the phenomenon under study. This indicates that long periods of fieldwork, traditionally considered to belong to ethnography, still play a role in the construction of ethnographic knowledge.

Although online ethnographic research is in many ways different from the more traditional ethnography based on face-to-face interaction, we pondered what significance this distinction ultimately has today. Presently, ethnographic research is characterized by the intertwining of the real world and the virtual world. Research situations are encounters of human and non-human actors. Central to our own research was to define the field, build interaction and rapport between researchers and participants, and implement observations in an online environment. Ethical reflection was a vital part of all stages of our research.

A researcher conducting digital ethnography needs to have adequate technical expertise. In our research, this need paved the way for co-research as many young people guided the researchers during the interviews and shared their own expertise in and about digital worlds. In the future, digital ethnography among youth could benefit of new collaborative ways of doing research. This could also increase young people's opportunities to influence research and define their own ways of participating, which in turn could encourage them to participate and have their own voices better heard.

5. Acknowledgements

The authors would like to thank all young people who participated in this study for their valuable contribution. Project DEQUAL: Capturing Digital Social Inequality Young digi-natives' asymmetrical agencies within socio-technical imperatives and imaginaries is funded by the Academy of Finland (SA 330574).

References

Abidin, C. & de Seta, G. (2020). Private messages from the field: Confessions on Digital Ethnography and its discomforts. *Journal of Digital Social Research* 2 (1), 1–19.

ALLEA - All European Academies (2017). *The European Code of Conduct for Research Integrity*. Revised Edition. Berlin: All European Academies.

Atkinson, P., Coffey, A., Delamont, S., Lofland, L. & Lofland, J. (eds.) (2001). *Handbook of Ethnography*. London: Sage.

Bennett, S. J. & Maton, K. A. (2010). Beyond the 'digital natives' debate: Towards a more nuanced understanging of students' technology experiences. *Journal of Computer Assisted Learning* 26 (5), 321–331. DOI: 10.1111/j.1365-2729.2010.00360.x

Beuving, J. J. (2020). Ethnography's future in the big data era. *Information, Communication & Society* 23 (11), 1625–1639. DOI: 10.1080/1369118X.2019.1602664

Boellstorff, T. (2008/2015 with a new preface by the author). *Coming of Age in Second Life. An Anthropologist Explores the Virtually Human*. Princeton, Oxford: Princeton University Press.

boyd, d. (2007). Why Youth ♥ Social Network Sites: The Role of Networked Publics in Teenage Social Life. In D. Buckingham (ed.), *Youth, Identity, and Digital Media*. Cambridge: The MIT Press, 119–142.

Burrell, J. (2009). The Field Site as a Network: A Strategy for Locating Ethnographic Research. *Field methods* 21 (2), 181–199. https://doi.org/10.1177/1525822X08329699

Caliandro, A. (2018). Digital Methods for Ethnography: Analytical Concepts for Ethnographers Exploring Social Media Environments. *Journal of Contemporary Ethnography* 47 (5), 551–578.

Cerwonka, A. & Malkki, L. H. (2007). *Improvising Theory. Process and Temporality in Ethnographic Fieldwork*. Chicago, London: The University of Chicago Press.

Collins, S. G & Durington, M. S. (2014). Networked Anthropology. A Primer for Ethnographers. London, New York: Routledge.

Davies, C. A. (2002). *Reflexive Ethnography: A Guide to Researching Selves and Others*. New York: Routledge.

van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society* 12 (2), 197–208.

El Founti, E., Leino, I. & Pyyry, N. (2021). "Haluutsä mennä kulmakauppaan pööpöileen? Tää on sun maailma." – Psykomaantiede ja kaupungin kanssa oppiminen. *Alue ja ympäristö* 50 (1), 139–147.

Eriksson, S. & Tuuva-Hongisto, S. (2019). *Nuorisotyön digitalisaatio 2030 – "Meidän tulisi osata tarjota nuorille työkaluja maailmaan, jota me emme vielä itse tunne"*. Mikkeli: Kaakkois-Suomen ammattikorkeakoulu, Xamk tutkii 11.

Geertz, C. (1973). *Interpretation of Cultures. Selected Essays*. New York: Basic Books.

Gibson, K. (2022). Bridging the digital divide: Reflections on using WhatsApp instant messenger interviews in youth research. *Qualitative Research in Psychology* 19 (3), 611–631. DOI: 10.1080/14780887.2020.1751902

Góralska, M. (2020). Anthropology from Home. Advice on Digital Ethnography for the Pandemic Times. *Anthropology in Action* 27 (1), 46–52.

Hallet, R. E. & Barber, K. (2014). Ethnographic Research in a Cyber Era. *Journal of Contemporary Ethnography* 43 (3), 306–330. https://doi.org/10.1177/0891241613497749

Helsper, E. J. (2021). *The Digital Disconnect. The Social Causes and Consequences of Digital Inequalities.* London: Sage.

Hine, C. (2000). *Virtual Ethnography*. London: Sage.

Hine, C. (2015). *Ethnography for the Internet. Embedded, embodied and everyday.* London, New York: Routledge.

Horst, H. A., Hjorth, L. & Tacchi, J. (2012). Media Ethnography Revised: An Introduction. *Media International Australia* 145, 86–93.

Ito, M., Baumer S., Bittanti M., boyd, d., Cody, R., Herr-Stephenson, B., Horst, H. A., Lange, P. G., Mahendran, D., Martínez, K. Z., Pascoe, C. J., Perkel, D., Robinson, L., Sims, C. & Tripp, L. (2009). *Hanging out,*

messing around and geeking out: kids living and learning with new media. Cambridge: MIT Press.
Ito, M., Matsuda, M. & Okabe, D. (eds.) (2006). *Personal, portable, pedestrian. Mobile phones in Japanese Life*. Cambridge: MIT Press.
Korjonen-Kuusipuro, K. & Kuusisto-Arponen, A-K. (2019). Socio-material Belonging – Perspectives for the Intercultural Lives of Unaccompanied Refugee Minors in Finland. *Journal of Intercultural Studies* 40 (4), 363–382. DOI: 10.1080/07256868.2019.1628725
Koskinen-Koivisto, E. & Lehtovaara, T. (2020). Embodied Adventures. An experiment on doing and writing multisensory ethnography. In T. Lähdesmäki, E. Koskinen-Koivisto, V. Čeginskas & A-K. Koistinen (eds.), *Challenges and Solutions in Ethnographic Research*. London, New York: Routledge, 21–35.
Koskinen-Koivisto, E., Lähdesmäki, T. & Čeginskas, V. (2020). Introduction: Ethnography with a Twist. In T. Lähdesmäki, E. Koskinen-Koivisto, V. Čeginskas & A-K. Koistinen (eds.), *Challenges and Solutions in Ethnographic Research*. London, New York: Routledge, xx–xxix.
Kozinets, R.V. (2002). The Field Behind the Screen: Using Netnography for Marketing Research in Online Communities. *Journal of Marketing Research* 39 (1), 61–72.
Kozinets, R.V. (2009). *Netnography: Doing Ethnographic Research Online*. London: Sage.
Lehtiniemi, T. & Ruckenstein, M. (2019). The social imaginaries of data activism. *Big Data & Society* 6 (1), 1–12. https://doi.org/10.1177/2053951718821146
Licheterman, P. (2017). Interpretive reflexivity in ethnography. *Ethnography* 18 (1), 35–45. https://doi.org/10.1177/1466138115592418
Lugosi, P. & Quinton, S. (2018). More-than-human netnography. *Journal of Marketing Management* 34 (3–4), 287–313. https://doi.org/10.1080/0267257X.2018.1431303
Markham, A. (2021). *Digital Ethnography: More a mindset than a tool*. YouTube lecture. https://www.youtube.com/watch?v=ZBIzTOXqVnc
Miller, D. (2011). *Tales from Facebook*. Cambridge: Polity Press.
Murphy, E. & Dingwall, R. (2001). The Ethics of Ethnography. In P. Atkinson, A. Coffey. S. Delamont, J. Lofland & L. Lofland (eds.), *Handbook of Ethnography*. London: Sage, 339–351.
Nakamura, L. (2008). *Digitizing Race. Visual Cultures of the Internet*. Princeton: Princeton University Press.
Pink, S., Horst, H., Postill, J., Hjorth, L., Lewis, T. & Tacchi, J. (2016). *Digital Ethnography: Principles and Practice*. London: Sage.
Pitt, C., Bell, A., Boyd, B. S., Demmel, N. & Davis, K. (2021). Connected Learning, Collapsed Contexts: Examining Teens' Sociotechnical Ecosystems Through the Lens of Digital Badges. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Article no 354, 1–14. https://doi.org/10.1145/3411764.3445635
Postill, J. (2009). What is the point of media anthropology? *Social Anthropology* 17 (3), 334–337. https://doi.org/10.1111/j.1469-8676.2009.00079_1.x
Postill, J. & Pink, S. (2012). Social media ethnography: the digital researcher in a messy web. *Media International Australia* 145 (1), 123–134.
Prensky, M. (2001a). Digital natives, digital immigrants. *On the Horizon* 9 (5), 1–6.
Prensky, M. (2001b). Digital natives, digital immigrants, part II. Do they really think differently? *On the Horizon* 9 (6), 1–6.
Pyyry, N. (2019). From psychogeography to hanging-out-knowing: Situationist *dérive* in nonrepresentational urban research. *Area* 51 (2), 315–323.
Richardson, K. (2015). *An Anthropology of Robots and AI: Annihilation Anxiety and Machines.* New York: Routledge. https://doi.org/10.4324/9781315736426
Robards, B. & Lincoln, S. (2017). Uncovering Longitudinal Life Narratives: Scrolling back on Facebook. *Qualitative Research* 17 (6), 715–730.
Rokka, J. (2010). Netnographic inquiry and new translocal sites of the social. *International Journal of Consumer Studies* 34 (4), 381–387. https://doi.org/10.1111/j.1470-6431.2010.00877.x
Standlee, A. (2017). Digital Ethnography and Youth Culture: Methodological Techniques and Ethical Dilemmas. In I. Castro, M. Swauger & B. Harger (eds.). *Researching Kids and Teens: Methodological Issues, Strategies, and Innovations*. Bingley: Emerald Group Publishing Ltd, 325–348.
Sumiala, J. & Tikka, M. (2020). Digital ethnographers on the move – an unexpected proposal. *Journal of Digital Social Research* 2 (1), 39–55.
Thompson, A., Stringfellow, L., Maclean, M. & Nazzal, A. (2021). Ethical considerations and challenges for using digital ethnography to research vulnerable populations. *Journal of Business Research* 124, 676–683.
Thulin, E. &Vilhelmson, B. (2007). Mobiles everywhere: Youth, the mobile phone, and changes in everyday practice. *Young* 15 (3), 235–253. DOI:10.1177/110330880701500302
Turkle, S. (1995). *Life on the screen. Identity in the age of the Internet*. New York: Simon & Schuster.
Weber, G.W. (2015). Virtual Anthropology. *Yearbook of Physical Anthropology* 156 (S59), 22–42.

Webster, A., Svalastog, A. & Allgaier, J. (2020). Mapping new digital landscapes. *Information, Communication & Society* 23 (8), 1100–1105.

Ecofeminism with *Onlife* Potential: Netnographic Approach to Secondhand Clothing Entrepreneurships in Instagram

Wendy Marilú Sánchez Casanova
Consejo Nacional de Ciencia y Tecnología
E-mail: wsanchez@enesmorelia.unam.mx

Abstract

Background and Context: Secondhand fashion does not only comprise small-scale purchase and sale but represents an *onlife* movement that begins in the virtual space, and then is moved into the real space, going back-and-forth in a continuous way as a strategy of aware consumption. This is explained by ecofeminism, that refers to "the potential of women to achieve an ecological revolution" (Warren 2004, 63) in which entrepreneurs have the task of recovering unique garments that are presented as attractive for their quality, low in cost, and that decrease the ecological footprint through their reuse. *Objective:* To explore the *onlife* potential of Mexican secondhand clothing entrepreneurships driven via Instagram, focused on acquiring used clothing as a sustainable and ecofeminist practice. *Methodology:* Netnography emerges as a group of techniques and resources for this *onlife* research (see Del Fresno 2011), and in this sense, participant observation prevails, that is intended on its own and as a netnographic technique (see Kozinets 2010). *Results and Discussion:* The ideological and operative basis of secondhand clothing entrepreneurships can be initially analyzed in the virtual space of Instagram (marketing, sensitization, and dialogue), and then this analysis can be continued in real spaces (deliveries, shipments, street markets, marches). So, cyber/ecofeminism is an opportunity for inquiry of consumption, from fast fashion oriented to aware, by way of secondhand fashion. *Conclusions:* In order to understand contemporary *onlife* situations, it is important to question the idea of human being and technology as separate entities, because it is obvious that devices and virtuality are already assimilable to certain emotional and physical aspects of persons, since they develop their lives through these sources. However, this is done at different levels. Finally, netnography offers diverse ways for communication, requesting the researchers to share their works throughout lots of real and virtual resources, available in the *onlife* scenario.

Keywords: feminism, sustainability, fast fashion, slow fashion, netnography.

1. Introduction

This work aims to explore the *onlife* potential of Mexican secondhand clothing entrepreneurships in Instagram, driven primarily by young women who are objecting the negative impacts of fast fashion as overproducer of collections to maintain the desire to purchase poor quality items to be disposed too soon.

Therefore, secondhand fashion is an action oriented to acquire used clothing as a sustainable and ecofeminist practice in which entrepreneurs have the task of recovering unique garments that are presented as attractive for their quality, low cost, and that decrease the ecological footprint through reusing. This practice starts in Instagram in terms of self-introduction to incentive purchases, and then is transferred to material places to complete all transactions.

Consequently, the discursive and functioning core of secondhand clothing entrepreneurships can be initially explored in the cyberspace, and then in the real space, although it is possible to go back-and-forth in a continuous way. This is an *onlife* performance of contemporary lifestyle, and it is the main reason for this research that is guided by the following question: how does that *onlife* performance unfold in secondhand fashion entrepreneurships?

In terms of the above, the current study investigates which practices make the initiatives of secondhand fashion to be referred within an *onlife* structure, or as a "blurring of the distinction between reality and virtuality" (Floridi 2015, 7). In this sense, netnography was applied via participant observation since I was a customer of some secondhand clothing entrepreneurships.

It should be noted that this research was inspired by my personal commitment to slow fashion, so some findings are reported with an autoethnographic bias (for this, I will be using the singular

first-person), which could be a limitation, so I assume the responsibility for a unique narrative until complementary research techniques are implemented.

2. Background and Context

Fast fashion is defined as "an accelerated business model that drives people to purchase more clothes motivated by low prices and multiple micro-seasons every year" (Hernández 2020, 9). This implies an overproduction of garments that negatively impacts the environment and the economy, as well as people's bodies, due to the imposition of beauty ideals focused on certain sizes, or unfair working conditions, especially in clothing factories.

The slow fashion movement questions the aforementioned aspects of fast fashion and offers new practices to reduce or avoid the copious supply of poor-quality clothing designed for planned obsolescence. Among its options is secondhand fashion, which focuses on acquiring used clothing. This kind of entrepreneurship, which is mainly sponsored by young women inspired by their commitment to self-esteem, self-care, and environment, set the task of recovering unique garments that are presented as attractive for their quality, low cost, and that decrease the ecological footprint by reuse.

This practice starts in social networks such as Instagram, by way of purchase or sale of secondhand garments, or other models of social and solidarity-based economy such as barter, or the visualization of local creations. This is accompanied by discussions about sustainability that is visualized as possible when use value prevails over exchange value, which can be characterized as a component of an ecofeminist strategy oriented to responsible consumption. Thus, secondhand fashion does not only comprise small-scale purchase and sale, but represents a social movement *onlife*, due to the oscillations between real and virtual backgrounds.

The cluster of women who have made secondhand fashion a way of life seems to respond to the invitation extended by Bina Agarwal (2004, 250):

> *I would call to struggle for resources and meanings. I would involve struggling with the dominant groups that have the ownership, power, and privilege to control resources, and these or other groups that control thinking about them through the media and educational, religious, and legal institutions.*

The supporters of slow fashion parade on the runway of environmental awareness and the breaking of paradigms that have long impacted the bodies of people and the body that is the planet, in the face of the repetitive practice of "the commodification of everything" (Wallerstein 1983, 16). In this sense, the offer of secondhand garments goes together with conversations about the need to change the preponderant scheme of consumption without awareness. When this discourse is combined with participating directly in events defined as feminist (street markets, marches, among others), it leads to ecofeminism, a term that refers to the "potential of women to fulfill an ecological revolution" (Warren 2004, 63).

Although ecofeminism has different facets, the common point consists in the "accusation of the consumerist and devastating alienation of the Earth and looks for creating a new model of human development or good-living" (Puleo 2019, 21). In this way, ecofeminism pays special attention to the interconnection of ecosystems, and the body is one of them.

This is against the perspective of fast fashion in which people's bodies are affected by planned obsolescence, low self-esteem, labor exploitation, and environmental pollution. Secondhand fashion is an alternative vision that challenges these problems, with the understanding that slow fashion is

not just consumption, but a lifestyle resulting from an ecological awareness with a gender perspective.

Much of the ecofeminist content of secondhand clothing entrepreneurships is displayed on the Web, so it is possible to identify cyberactivism, as an:

> *Strategy for establishing temporary coalitions of people who, using tools of that network, generate sufficient critical mass of information and debate so that this debate transcends the blogosphere and takes to the streets or distinctively modifies the behavior of many people (De Ugarte 2007, 122).*

Because of this growing potential, "Kaarina Kailo has set up the term 'cyber/ecofeminism'. With it, she seeks to overcome a certain orthodox view that understands ecological concern as something absolutely opposed to the technological revolution that has enabled openness to the virtual world" (Puleo 2019, 82).

In this way, secondhand clothing entrepreneurships are continuously moving between virtual and real spaces, which Floridi (2015, 7) characterizes as *onlife* in accordance with the following expressions:

1. the blurring of the distinction between reality and virtuality;
2. the blurring of the distinctions between human, machine, and nature;
3. the reversal from information scarcity to information abundance; and
4. the shift from the primacy of entities to the primacy of interactions.

The success of those entrepreneurships is supported by a convenient back-and-forth between the virtual and the real stages, assuming the devices as a crucial part of the body, thoughts, and emotions, in the space of social networks from which a subject, individual or collective, creates a profile as introduction. Then he/she/it offers information to obtain reactions and customers that contribute to her/his/its way of being situated in the world.

3. Methodology

This work is led by netnography that refers to "the researcher's attempt to acknowledge the importance of computer-mediated communications in the lives of culture members, to include in their data collection strategies the triangulation between various online and offline sources of cultural understanding" (Kozinets 2010, 60). The work involves strategies for analyzing co-presences (Pink et al. 2019), following one or more of the following paths: a) in the virtual, or the action and impact of individual and collective subjects through the multiple and available venues on the Web; b) from the virtual, that means the functioning and structure of cyberspace in terms of its capacity to generate social phenomena; and c) through the virtual, linked to impacts of virtual existence on real existence and inversely (see Hine 2004).

Therefore, an *onlife* approach can reduce the expected risks while researching on the Internet. However, it is necessary to indicate that not all studies can take this method, since "netnography is presented as a new discipline, or as an antidiscipline, or an interdiscipline, or a theory under-construction that is being developed to understand the social reality that is taking place in cyberspace" (Del Fresno 2011, 59). Meanwhile, Bárcenas & Preza (2019, 140) argue that the researcher needs "diverse actions and mechanisms allowing to move through the space, to trace an own route and relationships with the subjects that interact within it". With this the authors refer to the mixture between online and offline sources mentioned above.

At this point, participant observation prevails in this study, since it is considered the most important element of ethnography and involves "taking an actual role within the group or institution,

and contributing to its interests, or function, while directly experiencing those effects inside others" (Woods 2015, 49). In this way, information has been gathered from my own perspective as a customer of different secondhand clothing woman entrepreneurs on Instagram, that has been my primary fieldsite. Subsequently, participation in face-to-face street markets as well as bartering events were carried out. This data collection from different sales processes, and participation in events cited above, started in July 2020, and for this work there has been a preliminary end in April 2022.

Participant observation is also one of the possible techniques within netnography that could go far beyond a "participant-observational research based in online fieldwork" (Kozinets 2010, 60). As specified by Del Fresno (2011), netnography will not always involve the direct participation of the researcher, because it is important to focus on aspects such as the production of discourses, interchange of meanings, and contextualization, which not only requires the back-and-forth between the real and the virtual, but also between the researcher's overview and that of other people.

For all those considerations, netnography can provide "more authors, more perspectives, more readings, more creative and interpretative potential. This new perspective on ethnography is known as hypertextuality" (Ruiz 2008, 121–122). In this sense, the hypertext has been a primary source for data collection, specifically in Instagram, to generate an assessment from an external point of view to fulfill an indirect ethnographic role.

This switching back-and-forth between inorganic and organic spaces leads us to emphasize that "the cyberspace reinforces the holistic concept that ethnography has always owned" (Ruiz 2008, 122), an open outlook that is necessary for the understanding of *onlife* dynamics.

4. Results and Discussion

According to the netnographic analysis, hypertextuality is crucial to launch merchandise itself. This introduction is the moment in which certain Instagram secondhand clothing entrepreneurship profiles provide information and sensitization posts regarding the option for the so-called 'preloved' garments (see Figure 1). This means that these garments are referred to as liked and used by someone before the person who can acquire them at this moment.

This aware consumption focuses on the (re)appropriation of bodies. The performance of secondhand clothing entrepreneurships implies a thoughtful selection of the garments to purchase, and since they are unique and no exchanges or refunds are allowed, issues such as knowing one's own measurements and basic information of fashion styles are decisive. In addition, customers find themselves warmly treated by the entrepreneur through terms such as 'beautiful', 'pretty', or 'gorgeous', and this is not seen only as a strategy to get the loyalty of buyers, but as conviction for a view that every woman is important and valuable, no matter what her physical appearance (see Figure 2). This particularity is what gives secondhand clothing entrepreneurs the nickname of 'nenis' ('cute girls'), underlining one of the feminist slogans, related to care and attention among women.

Figure 1: Reasons to Get Preloved Clothing[28] (Renovarte Wardrobe Bazar 2021a).

Figure 2: Love Your Body[29] (Renovarte Wardrobe Bazar 2021b).

[28] Text inside image: "Reasons to Get Preloved Clothing. 1. I save money. 2. I find divine and good quality garments. 3. I support sustainable fashion. 4. I contribute to the recovery of local economy. 5.- I can always find brands at affordable prices". Text below image: "A few little motivations for getting preloved clothes".
[29] Text inside image: "Love Your Body #AwareFashion #SustainableFashion". Text below image: "Take care and love your body. In Renovarte Wardrobe Bazar we ensure to have unique sizes and garments for unique women".

In this sense, the hypertextuality is the catalyst for an *onlife* movement which is present in the three paths distinguished by Hine (2004): a) in the virtual, because secondhand clothing entrepreneurships display in the virtual to introduce their discourses and garments; b) from the virtual, due to entrepreneurs trust on virtual resources to generate contents for dialogue and marketing, such as posts or reels;[30] and c) through the virtual, although the first contact is usually in the virtual space, the delivery or shipment of garments, as well as certain events, such as street markets or marches, take place in real spaces.

Even though the functioning of the entrepreneurships starts from Instagram, a virtual space, it needs to be completed in the real plane enabling an *onlife* experience under certain philosophy. This philosophy can be ecofeminist (or feminist in general) through showing a bit of personal life linked to that feminist trend such as certain experiences, opinion on national or international events related to women and/or fashion, as well as activist participation. As a result, the ideological and functional basis of secondhand clothing entrepreneurships can be initially analyzed in the virtual space, and then moved into the real space, being possible to go back-and-forth in a continuous way.

Moreover, I emphasize that my presence in secondhand fashion has been as a customer, or follower, and not as an entrepreneur, so from the participant observation I am systematizing the information from the point of view of only one of the actors involved.

Reinforcing an *onlife* environment, I have carried on to participant observation in virtual and real spaces, since the functioning of secondhand clothing entrepreneurships is performed as follows:

1. Photography of garments are posted on Instagram profiles of each entrepreneurship, with their descriptions and prices.
2. Interested people comment on the post and the entrepreneur contacts the first on the list, who has a deadline to confirm the purchase by paying the amount into a bank account or setting up an appointment for the material delivery of the garment.
3. In the case of customers who do not live in the same city, an extra amount is paid for shipping. Even in the same city, a local courier service could be used for deliveries.
4. The appointments for material deliveries take place in crowded places for security reasons. The meeting is reduced to the delivery of the garment and its payment, if not previously paid.

It is important to say that in the virtual background there is always an initial risk of getting false or non-verified information. In this research, this risk has been assumed when shopping in a certain entrepreneurship for the first time. However, the slow fashion community is sustained by recommendations, either from an entrepreneur to each other, or between entrepreneurs and customers. Therefore, I have had the opportunity to interact with entrepreneurships recommended by well-known entrepreneurs, which for me is the equivalent of the technique known as 'snowball', by which one informant refers to another in order to add narratives. This technique "does not grow in undetermined directions, or randomly, but along the rails of social structures previously existing at the time of the researcher's arrival" (Rodríguez 1999, 94).

In this universe of entrepreneurships, mentions are very valuable because it is a way to get more customers. Therefore, the fact that the posts of determined entrepreneurship are replicated in the 'stories'[31] of another entrepreneurship is very appreciated. On this aspect, there is a debate about the irresponsibility of 'mention for mention', consisting of entrepreneurships introducing others as a simple exchange of favors. For this, it is proposed that the recommendations be based on a true knowledge of the reputation of the entrepreneur in question. This underlines the idea of aware

[30] A reel is a short video created from different tools offered by Instagram itself.
[31] Audiovisual content that is available in social networks as Instagram or Facebook only for 24 hours.

consumption through former information, needed for accurate decision making when acquiring goods and services, which then contributes to the ideological and functional basis of secondhand clothing entrepreneurships.

All the aforementioned features refer to the social and solidarity-based economy, that means a community perspective focused on enhancing a better development of human relations, under the premise that "consumption is the set of sociocultural processes in which the appropriation and uses of products take place" (García 2009, 58–59). *Onlife* scenario, that is, the oscillation between real and virtual backgrounds, facilitates the consolidation of sales, trust, as well as the permanence of secondhand clothing entrepreneurships.

As described above, these entrepreneurships can be characterized as ecofeminist (this is my perspective and not precisely that of the entrepreneurs, as some define themselves as feminists in general), starting with the noticeable and reiterated disapproval of the prevailing model of consumption through posts that drive an awareness in line with those related to the sale of garments; the praising of self-esteem and self-care of the individual body; as well as the underlining of the importance of the body that we share with other species, the Earth, including tips for fixing garments and optimizing fashion styles without accepting the constraints of fast fashion (see Figure 3).

It should be noticed that this work is related to my personal commitment to slow fashion, so the findings have certain autoethnographic nuances (and therefore I am using the singular first-person). Nevertheless, I do not consider unethical the fact that a researcher takes the role of a customer as an approach to this field of study, due to personal interest. What should prevail in any case is that the study subjects may know who they are, to which I would add their right to access to the results, and its corresponding discussion.

Figure 3: Clothing Fixing[32] (Calíope Bazar 2021).

[32] Text below image: "Sometimes clothes need healing, nothing that thread, needle, and nice buttons cannot fix".

5. Conclusions

Cybergeography makes achievable the inclusion into the public sphere of people, groups, and movements that used to be ignored or unknown. This is possible as the capacities for spreading are manifested in a multiplicity of virtual resources available to those who can be 'connected'. In this sense, cyber/ecofeminism is one of many contemporary opportunities to enhance *onlife* structures.

As a result, the *onlife* path is both theory and methodology. In the first sense, it is an approach of discussion about how current social relations develop, in which very little is done without virtual elements. As for the second sense, if individuals and collectivities move between the real and the virtual backgrounds, then it is important to approach this field of study with suitable research techniques, which is why netnography has broadened its horizons by dealing with the strictly virtual, or the influence of the virtual on the real, to the point of defining a social display within the framework of *onlife*.

For research focused on the *onlife*, it is necessary to go beyond adding the prefix 'net' to ethnography, which implies going "against the gloomy theories of alienation that have been fashionable for a long time, it now becomes possible to investigate in more detail how technologies actually help to shape new relations between humans and world" (Verbeek 2015, 219). This entails eradicating the idea of human being and technology as separate entities that merely converge in specific situations, when it is obvious that devices and virtuality are already assimilable to certain emotional and physical aspects of persons, because they develop their lives through these resources, in different levels.

Finally, it is worth noting the heterodoxy of netnography, which "demands attending to alternative ways of communicating" (Pink et al. 2019, 30). In this context, I decided to create an Instagram account (Moda Cíclica - @moda.ciclica.inv; see Figure 4) aimed to share progresses and results of my research within the community of secondhand clothing entrepreneurs.

Figure 4: Clothing Exchange[33] (Moda Cíclica 2022).

[33] Text below image: "Today I took part in the 8th edition of @2vsegundavuelta (clothing exchange project in Morelia, Mexico), and as always, I got treasures that will write a new story".

This opening of spaces for collecting of information, especially that which can be obtained from the Web, invites us to rethink practical concerns of social research, in which the current trend claims that there are no objects, but subjects of study. For this, an inherent issue is the discussion on what extent is the subject a subject and not just a means to achieve a result that translates into articles for a reduced academic sector. After all, the ethics of researching (in all disciplines) is its commitment to be distributed in accessible ways to different audiences, and for this we have lots of resources, real and virtual, which makes the process of this distribution an onlife scenario in itself.

6. References

Agarwal, B. (2004). El debate sobre género y medio ambiente: lecciones de la India. In V. Vázquez García & M. Velázquez Gutiérrez (coord.), *Miradas al futuro. Hacia la construcción de sociedades sustentables con equidad de género*. Mexico City: PUEG/CRIM/CP, 239–285.

Bárcenas Barajas, K. & Preza Carreño, N. (2019). Desafíos de la etnografía digital en el trabajo de campo onlife. *Virtualis* 10 (18), 134–151.

Calíope Bazar [@caliopeelbazar] (2021). *A veces las ropitas requieren curación, nada que hilo, aguja y botones bonitos no puedan solucionar* [Photography]. Instagram, January 31st, 2021. https://www.instagram.com/p/CKuN_HBh4M2/?igshid=YmMyMTA2M2Y=

Del Fresno, M. (2011). *Netnografía. Investigación, análisis e intervención social online*. Barcelona: UOC.

De Ugarte, D. (2007). *El poder de las redes. Manual ilustrado para personas, colectivos y empresas abocados al ciberperiodismo*. Madrid: Ediciones El Cobre.

Floridi, L. (2015). The Onlife Initiative. In L. Floridi (ed.), *The Onlife Manifesto. Being Human in a Hyperconnected Era*. Oxford: Springer Open, 7–13.

García Canclini, N. (2009). *Consumidores y ciudadanos. Conflictos multiculturales de la globalización*. Mexico City: Ramdom House Mondadori.

Hernández, C. (2020). Moda rápida: la industria que desviste al planeta. *¿Cómo ves?* 22 (57), 6–12.

Hine, C. (2004). *Etnografía virtual*. Barcelona: UOC.

Kozinets, R. V. (2010). *Netnography. Doing Ethnographic Research Online*. Los Angeles: Sage.

Moda Cíclica [@moda.ciclica.inv] (2022). *Hoy participé en la 8a edición de @2vsegunda vuelta y como siempre salí con tesoros que escribirán una nueva historia* [Photography]. Instagram, March 05th, 2022. https://www.instagram.com/p/CavUSV4PzU3/?igshid=YmMyMTA2M2Y=

Pink, S., Horst, H., Postill, J., Hjorth, L., Lewis, T. & Tacchi, J. (2019). *Etnografía digital. Principios y práctica*. Madrid: Morata.

Puleo, A. H. (2019). *Ecofeminismo para otro mundo posible*. 7th ed. Madrid: Ediciones Cátedra.

Renovarte Wardrobe Bazar [@renovartewardrobe_bazar] (2021a). *Unos cuantos motivillos para comprar prendas Preloved* [Photography]. Instagram, January 31st, 2021. https://www.instagram.com/p/CKu0f82HMrU/?igshid=YmMyMTA2M2Y=

Renovarte Wardrobe Bazar [@renovartewardrobe_bazar] (2021b). *Cuida y ama tu cuerpo, en Renovarte Wardrobe Bazar nos encargamos de tener tallas y prendas únicas para mujeres únicas* [Photography]. Instagram, February 10th, 2021. https://www.instagram.com/p/CLHj3sYHCB4/?igshid=YmMyMTA2M2Y=

Rodríguez Bilella, P. D. (1999). Evaluación de proyectos y triangulación: Acercamiento metodológico hacia el enfoque centrado en el actor. In A. Ocampo (ed.), *Memoria del Segundo Taller Electrónico sobre Evaluación de Proyectos de Reducción de la Pobreza Rural en América Latina y el Caribe*. November 2nd–December 10th 1998. San José: Instituto Interamericano de Cooperación para la Agricultura (IICA), 91–100.

Ruiz Torres, M. A. (2008). Ciberetnografía: comunidad y territorio en el entorno virtual. In E. Ardèvol, A. Estalella & D. Domínguez (coord.), *La mediación tecnológica en la práctica etnográfica*. Donostia: Ankulegi, 117–132.

Verbeek, P. P. (2015). Designing the Public Sphere: Information Technologies and the Politics of Mediation. In L. Floridi (ed.), *The Onlife Manifesto. Being Human in a Hyperconnected Era*. Oxford: Springer Open, 217–227.

Wallerstein, I. (1983). *Historical Capitalism*. London: Verso Editions.

Warren, K. J. (2004). Feminismo ecologista. In V. Vázquez García & M. Velázquez Gutiérrez (coord.), *Miradas*

al futuro. Hacia la construcción de sociedades sustentables con equidad de género. Mexico City: PUEG/CRIM/CP, 63–70.
Woods, P. (2015). *La escuela por dentro. La etnografía en la investigación educativa*. Barcelona: Paidós.

Digital Remediation and Visual Manipulation: Blogs as Breathing Spaces for Chinese Tattoo Wearers and Enthusiasts

Songqing Li

Department of Applied Linguistics, Xi'an Jiaotong-Liverpool University

email: songqing.li@xjtlu.edu.cn

Abstract

In the post-digital era, digital remediation plays a pivotal role in delivering political messages. This study examines postings of corporeally mediated tattoos in a China-based blog, focusing on in what sense the blog provides a breathing or even emancipating space for Chinese tattoo wearers and enthusiasts to control the viewer's perceptions of tattooing and negotiate long-standing stigmatized associations of tattoos there. The study examines 305 postings collected from the blog covering a period between 04/24/2020 and 05/03/2021, combining quantitative and qualitative approaches. Methodologically, it analyses the distribution of the tattoos across body parts as displayed. By applying multimodal perspectives, this study also investigates photographic techniques harnessed and exploited in turning body narrative into digital narrative. Results of the study suggest that digital remediation facilitates personal expression of tattoo wearers and photographic techniques play a critical role in introducing ready alignment of the viewer with the postings. The study thus adds quantitative inquiries to existing, mostly qualitative, studies of tattoos which usually rest on interviews with tattoo wearers, enthusiasts and artists for an account of tattoo narratives in connection to personal expression and self-definition. Its findings are awfully inspiring to socially stigmatized and marginalized (sub)groups to circumvent social, cultural and political barriers to communicate and make their stories heard to more people.

Keywords: blog, tattoo, digital remediation, visual manipulation, self-expression, China

1. Introduction

For a long time, tattooing has been a subject of theoretical and practical interest in various fields, including anthropology, art history, classics, media and literacy studies, and sociology. It is not, however, until fairly recent interest in sociolinguistic studies of tattoos (Hiramoto, 2015; Koller & Bullo, 2019; Martin, 2019) and skinscape studies of tattoos within linguistic and semiotic landscapes research (Peck & Stroud 2015; Peck & Williams 2018; Roux, Peck & Banda 2019) that signs inscribed on the human body (i.e., inscription tattoos) no longer remain entirely neglected in applied linguistics and sociolinguistics. Both sociolinguistic and skinscape studies of tattoos, however, attend almost exclusively to inscription tattoos; let alone digitally *re*mediated tattoos resulted from the transmission or transformation of physical expressions of corporeally mediated tattoos into virtual expressions of digitally mediated ones. Digitally *re*mediated tattoos should have already become the data of analysis, considering an increasingly important role Internet-based social media are playing in (re)defining, commodifying, and alternating "the characteristics adopted by contemporary tattoo" (Walzer & Sanjurjo 2016, 73). Besides, Internet-based social media have already grown into research sites of sociolinguistics (Deumert 2014) and semiotic landscape (Ivković & Lotherington 2009; Ifukor 2011; Kallen, Dohnnacha & Wade 2020). This article contributes to addressing this gap by examining blog postings of corporeally mediated tattoos from a multimodal perspective, focusing specifically on the adaption and transformation of tattoos based on an understanding of visual manipulation in connection to digital remediation. One of the objectives is to address whether blogs provide tattoo wearers a breathing space for meaning negotiation, cultural rebellion, and empowerment establishment.

The majority of work on tattooing from its inception has been preoccupied with design, meanings, and purposes, especially in connection to processes such as identity, consumerism, and modernity (e.g., Lane 2014; Schildkrout 2004). Rosenblatt (1997, 309), for example, claims that tattoos simultaneously are signs of personal identity and compulsory markers of difference from

people at large. For discourse analysts Koller and Bullo (2019), tattoos as forms of non-verbal communication meet both an interpersonal and an experiential (or ideational) function in terms of social semiotics (Kress and van Leeuwen 2006). It is ideational because tattoos "allow people to express themselves and their experiences through their bodies" (Koller and Bullo 2019, 4). It is interpersonal in the sense that tattoos can be an important medium for conformity, resistance, and negotiation of established relations (Atkinson 2002), and demarcating collective belonging of a social group (DeMello 2000). The ideational and interpersonal functions of tattoos can be equally observed in digitally remediated tattoos displayed at the blog or, more precisely, in blog postings of corporeally mediated tattoos. Similar to corporeally mediated tattoos, digitally remediated tattoos could be taken as a reliable resource for examining narrative performance and meaning negotiation. This position cannot be more sensible and adequate if referring to "self-expression and community development" affordances of blogs (Miller & Shepherd 2004; see also Myers 2010; Page 2011; 2018; Jaworska 2018). Besides, as proved by previous research, blogs provide sites for marginalized individuals and communities to negotiate meanings, counter social prejudice, and establish a sense of empowerment (e.g. Trainer et al. 2016; Limatius 2020; 2019; Mallya & Susanti, 2021). Focusing exclusively on ideational functions of digitally remediated tattoos, several studies have shed light on how tattoo wearers rely upon blogs to express themselves and their experience through their bodies (Hiramoto, 2015; Koller and Bullo, 2019). Yet, with these exceptions, rarely have efforts been made to examine how blogs provide tattoo wearers a breathing space for meaning negotiation, cultural rebellion, and empowerment establishment.

In its focus on the postings of corporeally mediated tattoos at Tattoo Exchange Circle, a China-based blog where Chinese tattoo wearers and enthusiasts share and discuss different types of tattoos, this article discusses whether and how corporeally mediated tattoos are capitalized on as a new resource for engaging with the viewer and manipulating their perceptions of tattooing, as well as negotiating sociocultural meanings of tattooing in China. Specifically, the article explores semiotic means of digital remediation that probably make tattoos open to adaption and transformation for intended interpersonal meanings. In particular, it aims at illuminating how body narrative can take a different trajectory, a politicized one because of photo or visual manipulation, which normally goes undetected, on the viewer's interactions to/with post-remediated tattoos that are often cultivated with specific forms of affect in the mediascape (Wee, 2016). Two key aspects or questions in which engagement and persuasive manipulation emerge and get solidified are to be addressed: (1) in what sense does digital remediation facilitate body narrative in tattoos in the context of blog? and (2) what semiotic strategies are harnessed and exploited in the process of digital remediation? Notably, the term "manipulation" is not used here to be deceptive or misleading, but to aid the viewer in accessing tattoos previously inscribed on the skin, increase their perceptions of tattoos, and influence their attitudes and behaviors, with or without their awareness, by introducing significant limits to negotiability.

Out of consideration of the widely held position on tattooing both as an individual and as a cultural affair (Fisher 2002), a brief introduction to tattooing in China is in order to better contextualize the digitally remediated tattoos.

2. Tattooing in China

The history of tattooing in China is somewhat difficult to trace, despite the prevalence of traditional conceptions of tattoos in premodern China (see Lei 2009 for overview). The most famous tattoo in

Chinese history may come from the legend of the Chinese general Yue Fei[34] on whose back his mother tattooed four characters "Jin Zhong Bao Guo" (*lit.*, Serve his country with ultimate loyalty). But it was mainly marginalized social and cultural groups, or "the bottom rung of society" (Lei 2009, 104) that were likely to get tattooed. In premodern China, tattoos had been used as a penalty manner for a very long period. The stereotypical association of tattoos with gangsters, prisoners, and crime pervades even in China today. Notwithstanding this historically deep-rooted perception, tattoos are nevertheless being growingly embraced by Chinese younger generations, especially in the major cities of Shanghai, Guangzhou and Shenzhen, with the desire for self-expression of individuality as the chief drive for tattoo acquisition in the modernizing Chinese society (Ang 2019; Davey & Zhao 2019). The creeping popularity of tattoos with Chinese young people, as observed by Ang (2019), does not deny the long-standing stigmatized associations of tattooing there; nor does it declare that these prejudices no longer persist in many parts of Chinese society, particularly among the elderly. These facts cogently account for why tattoo wearers tend to hide signs with long sleeves or others, and why they are, whether officially or not, excluded from a range of occupations including civil service, army, police, and education.

The foregoing account of the sociocultural practices and conventional views of tattooing in China is certainly supportive of the argument that the blog likely provides a site and a channel, by means of which stigmatized Chinese tattoo wearers circumvent social, cultural and political barriers to communicate and make their stories heard to more people. Differently stated, the blog in China might be deployed as a breathing or even emancipating space for multifarious functions of tattoos, including personal expression, self-definition, and cultural rebellion.

In line with the research questions to be addressed, the objective of this article, however, is not to examine what particular self-identities are presented or what type of a virtual community is constructed by postings of bodily mediated tattoos. This article does not aim to discuss the specific impact of postings resulted from photo manipulation either. Rather, it is semiotic strategies adopted and implemented during the digital remediation process for photo manipulation that my investigation into postings focuses on. Surprisingly or not, such strategies have hardly been the consideration of existing, mostly qualitative, studies which usually rest almost exclusively upon ideas and attitudes of tattoo wearers, enthusiasts, and artists for an account of tattoo narratives.

3. Digital remediation: Turning body narrative into digital narrative

The term "remediation" comes from the work of Bolter and Grusin (2000), but it originates with the idea put forth by Marchall McLuhan in his book *Understanding Media* that "the 'content' of any medium is always another medium" (McLuhan 1964, 8). This term has been used in many different ways (see Prior and Hengst 2010 for overview). Used in this article, the notion of remediation refers exclusively to the adaption and transformation of the original work in one medium into another within "transmedia chains", i.e., the sequencing of semiotic resources transmitted or transformed into a different medium, and through "transmedia traces", i.e., the way a preceding transmission or transformation in a different medium is indexed and thematized in another medium (cf. Androutsopoulos 2021).

A number of empirical studies of tattooing have already given substantive evidence to the conceptualization of inscription tattoos as the consequence of semiotic remediation (e.g., Alvarez

[34]Yue Fei served the South Song Dynasty. During the battle with northern enemies the Field Marshall under whom Yue Fei served betrayed the South Song and went over to the enemy. In protest Yue Fei resigned and returned home. His mother grew angry with him, telling him that his duty was first and foremost to his country, despite all else.

2020; Bengtsson, Ostberg & Kjeldgaard 2005; Koller & Bullo 2019; Patterson 2018; Roux, Peck & Banda, 2019). Aside from exemplifying inscription tattoos as products of semiotic remediation, these studies also make salient individual agency in tattooing to artistically set themselves apart through their body from constraints of their surrounding society for various functions. Hiramoto's (2015) study of tattoos worn by the immigrant Japanese in Hawaii is a case in point. Hiramoto directed our attention to the intertextual appropriation of texts and visual images performed in tattooing for identity display. Specifically, with the medium of tattooing, the Japanese language and associated traditional images are imbued with additional indexicals, which distances the immigrant Japanese from native Japanese, and have increasingly gained local-specific cultural values. Hiramoto analysed postings collected from tattoo forums and blogs; however, her analysis is confined to the transmission of signs through the embodied medium of the skin. An exclusive focus of analysis on inscription tattoos alone brings to light the limited or narrow spectrum of media actualized in the course of remediation; it also overlooks the fact that the postings as displayed result from digital remediation and the transformative impact of this process.

That this is a problem becomes clear when considering the possibilities held by postings not only for transmission of signs through the embodied medium of the skin, but also most of the times for adaption and transformation through digital media in capture of inscription tattoos and post-capture photographic montage like cropping. Alongside access to "polymedia" and "polymedia repertoire" (Androutsopoulos 2021; Tagg & Lyons 2021), transmedia chains enable "refashioning of materials and practices" and the creative "borrowing and adapting [of] materials and techniques whenever possible" (Bolter and Grusin 2000, 68). Prior and Hengst (2010) point out a direct correlation between remediation and "repurposing", although remediation does not always necessarily entail repurposing. Viewed as recycling of materials or content from one medium in another medium for a different purpose, the idea of repurposing, however, is not just to replicate an earlier form, but also to exploit new meanings that claim "to offer an experience that the other forms cannot" (Bolter and Grusin 2000, 68). Related to postings in question, a new experience of the viewer stemmed from digital remediation can be enriched specifically in terms of affect. The landscape, as Jaworski and Thurlow (2010, 4) point out, is a "place of affect". So is the mediascape--- "the power of many forms of media lies not so much in their ideological effects, but in their ability to create affective resonances independent of content or meaning" (Shouse 2005). As my discussion shall show below, photographic techniques harnessed and exploited in digital remediation processes are specific media "practice" (Couldry 2012) through which specific affective formations take form and an affective relationship between the viewer and the posting is built, an important step in coming to terms with tattooing. This step, according to Ahmed's (2004) conceptualization of affect, is necessary for visual manipulation in operation and subsequent negotiations of sociocultural meanings of tattooing.

The working of repurposing via digital remediation, emphatically, is contingent substantially upon photographers as agents rather than blogs and postings alone. There are two main reasons. Firstly, throughout the remediation processes, the transformation may be more or less palpable, more or less significant, but it always showcases varying degrees of agency, creativity and innovation in many aspects. Secondly, the embodied self-expression via tattoos is not necessarily the self-narrative; the self-narrative is in the tattoo but reliant upon the viewer, a point indicating the essential connection of how tattoos are captured and displayed for the viewer's interaction, engagement and perception.

This argument cannot be more apparent by virtue of the changing status of photography: photography has lost its special status as a visual medium to provide factual evidence of a human

activity and as a tool for an individual's identity formation and communication, but has increasingly developed into a means of visual manipulation (Djonov, Tseng 6 Lim 2021; Bateman 2021). On the other hand, since photography is a highly malleable and thus processable medium, the self-efficacy of visual literacy as a weapon against visual manipulation by photography is often called into question (Emme & Kirova 2005; Messaris 1994; 2012). At this moment it is worth noting that photo manipulation as human action that influences the appearance of a photographic image ranges from necessary pre-capture decisions to post-capture changes, all stages of which influence the implied meaning of a photographic image. Yet, photo manipulation is frequently discussed by focusing on the post-processing component independent of the pre-capture and capture steps, although the influence of capture cannot be ruled out. In fact, through use of focus, exposure and choice of shot, the camera can offer different versions of an inscription tattoo. It becomes presumable to render photographic techniques harnessed and exploited in capture as connected intimately to ramifications photo manipulation may have to the viewer's engagement with, and perceptions of, postings. In this article, the examination into photo manipulation of postings rests on the premise that postings are the direct products of capture in line with pre-capture decisions of what photographic techniques are to be harnessed and exploited.

4. Data and methods

The objective of this article, as noted above, is to examine the pictures of corporeally mediated tattoos posted at Tattoo Exchange Circle (hereafter TEC) for discussing semiotic strategies adopted and implemented during the digital remediation process for photo manipulation. Founded on 23 March 2020 and affiliated to Zhihu[35], TEC is dedicated to providing a virtual space for Chinese tattoo wearers and enthusiasts to share and discuss pictures of various (types of) tattoos (https://www.zhihu.com/club/1225447291425058816). The data for this study consist of all the postings displayed prior to 3 May 2021 but exclude excerpts from the blog and profiles of the bloggers. A corpus of 305 postings, which present or depict a number of participants, such as animals and real and fictional characters, and objects including artwork, letters, landscapes, and shapes (e.g., stars, flowers, hearts), were finally assembled.

This study takes a mixed method for addressing the aforementioned research questions. A preliminary quantitative examination was first conducted to discover the body parts from which the pictures of inscription tattoos were taken. This selectivity-related examination is largely associated with "different geosemiotics" (Scollon & Scollon, 2003) and varying degrees of visibility and, thereafter, the communicativity of inscription tattoos, as defined by a particular body part where a sign is inscribed (Peck & Stroud, 2015: 139). Related to pre-capture decisions on choices of photographic techniques to capture physically (in)visible signs on the skin, the quantitative analysis is able to address the question whether the blog affords a breathing or even emancipating space for self-expression by Chinese tattoo wearers. The postings were categorized in light of the visibility or invisibility of body parts. In this study the term "arm" refers to upper arm, biceps, elbow and forearm, precluding wrist where a sign is normally observable. In a posting where there are several adjacent body parts inscribed with more than one participant or object, it was counted only once in light of the main or most salient body part on which the signs were inscribed, taking into account the bodily orientations of the depicted participants or objects. For example, the posting, as shown in Figure 1,

[35] Zhihu is a social media platform like Baidu for information sharing in China. According to its third quarterly financial report released on 22 November 2021, it has monthly active users amounted to around 101 million by 30 September 2021 (https://www.chinaz.com/2021/1123/1332309.shtml).

was grouped into the category “arm”, despite the coverage of a tiger across the right shoulder and the right chest alongside a Buddha on the right upper arm. Where the location of a sign is hard to identify, they were labelled as “unsure”.

Figure 1: “Buddha” around upper arm and shoulder and “Tiger” on right ribcage

A series of quantitative analysis was further undertaken to discuss photographic techniques harnessed and exploited in turning body narrative into digital narrative. The quantitative analyses of photographic techniques are not independent of, nor isolated from, but complements, the first quantitative inquiry for a comprehensive picture of interactions between the posting and the viewer. As shown in Figure 1, photographic images are always presented from a particular perspective and position viewers within that perspective. With regard to postings, this is concerned almost exclusively with photographic techniques harnessed and exploited in digital remediation processes for engaging with viewers and manipulating their perceptions of the depiction. Since tattoos are signs comprising texts, visual images, and colors, in describing photographic techniques I drew upon a social semiotic approach (Kress & van Leeuwen 2006; Painter, Martin & Unsworth, 2012) to discuss a number of functionalist dimensions that images can serve in meaning making and visual manipulation. Specifically, multimodal perspectives suggest three categories of visual meaning systems— representational/ideational meaning, interactive/interpersonal meaning, and compositional meaning (Kress & van Leeuwen, 2006) — available to be mobilized for the encodement of social meaning into visual image relevant to social relationships between the viewer and the depicted person or object. As stated earlier, the purpose of this study is not to examine all these three categories, but to focus on interactive meanings that are especially significant to its thesis concerning issues of visual manipulation. The interactive meaning considers the ways that “contact”, “distance”, and “point of view” experienced by the viewer becomes part of a sign’s meaning. All of the aspects are indicators of photographic techniques deployable to invite the viewer to interact and engage with a depicted participant or object through which their attitudes towards the depiction can be visually manipulated.

In addition, modality in social semiotics could also be taken as a photographic technique for repurposing, because it "relates both to issues of representation … and to questions of social interaction, because the question of truth is also a social question" (van Leeuwen 2005, 160). Thus, four aspects of this social semiotic, "distance", "contact", "point of view", and "modality" are of particular relevance to the issues of photographic techniques and their correlation with repurposing. Alongside the quantitative findings of photographic techniques, a qualitative examination of each photographic technique will be conducted for an illustration of the visual manipulation of photographic techniques.

Here, it is important to note that Kress and van Leeuwen's (2006) idea of the visual expression of modality related to the realism of a depiction pays attention to modality markers including color, detail, light, shadow, and shade that can be hinged on to impose a view of truth and reality that is hard to counter. In the context of tattoos, the depiction of a participant or object, however, is often iconic and stylized, being relatively low in the degree of certainty. In other words, the articulation of a participant's or object's detail in postings as well as its corresponding degree of certainty is predetermined prior to very capture. For the quantitative analysis of modality, this study thus does not simply adopt Kress and van Leeuwen's (2006) idea of the visual expression of modality. Painter, Martin and Unsworth (2012, 31) deflect our attention away from degrees of certainty by redefining modality as a means of "realizing a system of reader alignment, or PATHOS, and as affording particular ways of presenting a character's emotions or affect". In children's picture books, Painter, Martin and Unsworth have identified three styles of character drawing—minimalist, generic, and naturalistic—on the basis of the degree of detail and realism of drawing, and the different kinds of alignment that are encouraged. The affectual dimension of modality is helpful to discuss particular ways affordable by modality to impact the viewer's emotions or affect. There are two more points worthy of our attention. Firstly, authenticity is usually an effect of semiotic authentication processes. Thus, it is "authenticity effect", rather than authenticity, achieved through the authenticating practices (Bucholtz 2003, 408) of bloggers who rely on semiotic means that is more accurately connected to the visual manipulation of the viewer's attention to the depicted. Secondly, the depiction of affect is not carried by detail alone, as modality markers usually "behave in relatively independent ways" with a high possibility, if not absoluteness, of being incongruous with each other for the intended truth and reality (Kress and van Leeuwen 2006, 171). This study thus intends not to examine all modality markers or semiotic means and the different impacts they can afford on visual manipulation, but merely to focus on the background's detail and lightness likely opted in capture for contrastive highlighting of a depicted participant or object in postings.

5. Findings

5.1. Distributions across body parts

The quantitative examination into the body parts from which the inscription tattoos were photographed for uploading and displaying at TEC, unsurprisingly, showcases an extremely uneven distribution. As suggested in Figure 2, listed at the top is arm with more than 50% of the total, which is followed by back, leg, chest, shoulder, and stomach amounting to 14.4%, 14.1%, 5.2%, 3.0%, and 2.6%, respectively. Differentiated from those on neck and nape, hand, and wrist, tattoos on the other body parts can well be defined as invisible when eminently covered with long sleeve shirts, pants or shoes. In sum, the invisible tattoos in light of placement amount to 95.8% of the total. Tattooing, as noted earlier, is still highly stigmatized in China today. The quantitative finding of the postings

regarding the predominant portion of invisible tattoos is in alignment with concerns of tattooed individuals about disfavor within this hegemonic order proscribing the bodily expression of tattoos. But the communicative value of the same tattoos is increased after being yielded visible at the blog, which demonstrates compellingly that the blog facilitates the self-expression affordance of tattoos. This, for sure, is preconditioned by the deliberate selection of inscription tattoos for digital remediation. This quantitative finding, additionally, suggests forcefully a focus on the invisible tattoos for a series of quantitative investigation into photographic techniques probably harnessed and exploited in digital remediation processes.

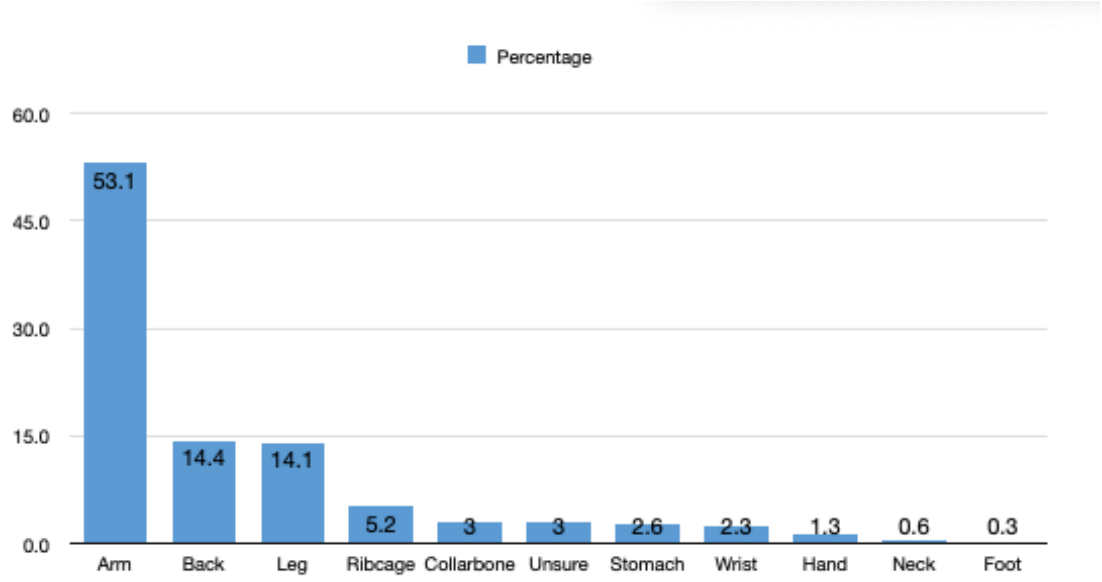


Figure 2: The distribution of tattoos on body parts

5.2. Photographic techniques

The focus of analysis below, as noted above, is moved on to the four aspects of photographic techniques—"distance", "contact", "point of view", and "modality", addressing whether and how they are harnessed and exploited in the capture of corporeally mediated tattoos to be displayed at the blog.

5.2.1. Distance

One of the most conspicuous observations from the postings is the tactful option of distance between the viewer and the depicted participants or objects. Distance is the association of physical proximity and intimacy. Kress and van Leeuwen (2006) have discussed social distance between the presented participant or object and the viewer often in terms of size or frame and shots opted in presenting or depicting a participant or object. For example, if shown in long shot and occupying about a quarter of the height of the portrait format frame, a participant or object is displayed for contemplation only, thereby the viewer being discouraged to interact with the presented participant or object.

Figure 3a

Figure 3b

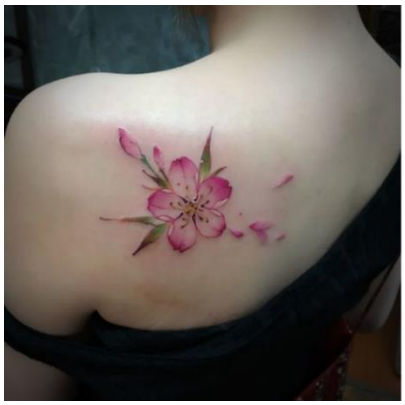


Figure 3c

Figure 3d

The digitally remediated tattoos of the data, however, are not always displayed in the portrait frame calling for a vertical format, but sometimes in the rectangle or square frame either calling for a horizontal or vertical format. Despite this, the same kind of distinctions can still be made with respect to the depiction of all the participants and objects. For example, the postings reproduced in Figures 3a, 3b, 3c and 3d differ by shape, size, and/or format; however, they are identical in that they are shown in whole and presented within the viewer's reach. The participants and objects in these cases are depicted at extremely close distance, close distance, middle distance, and long distance, respectively[36] Figure 4 presents the findings of extreme close-up, close-up, medium shot, and long shot opted in the depiction of all the participant and objects. As it shows, approximately 60% of the participants and objects are depicted either in extreme close-up or close-up, generating a personal distance at which they can be viewed in a specific environment and in a particular part of the body, getting the viewer deeply involved with the depicted. Even in the postings like Figure 3b often categorized into medium shot, we as viewers nevertheless feel a high degree of solidarity or closeness to the depicted. Thus, only 7% are presented in a long shot, creating maximal social distance whereby the viewer is detached from the tattoos.

[36] The continuum ranging from extremely close-up to long shot is primarily defined on the basis of the researcher's own perception.

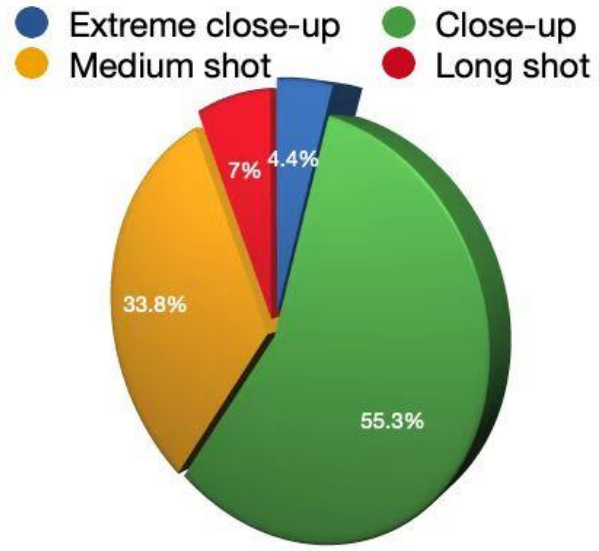


Figure 4: Distance and social intimacy here

5.2.2. Angle

Visual angle is known as an essential parameter of how the viewer sees the content of postings. Dependent on object perception, visual angle affects emotional responses to postings. Kress and van Leeuwen's (2006, 129–143) systems of involvement and power are further ways of positioning the viewer, and these depend on the use of perspective, which creates a "subjective" position by requiring a participant or object to be viewed from a particular angle. The majority of the signs on the skin seem to have been captured in the frontal and eye-level angles, so that the participants or objects in the photographic images are not just shown frontally (73.4%, see Figure 5) but level with the eyes of the viewer (70%, see Figure 6).

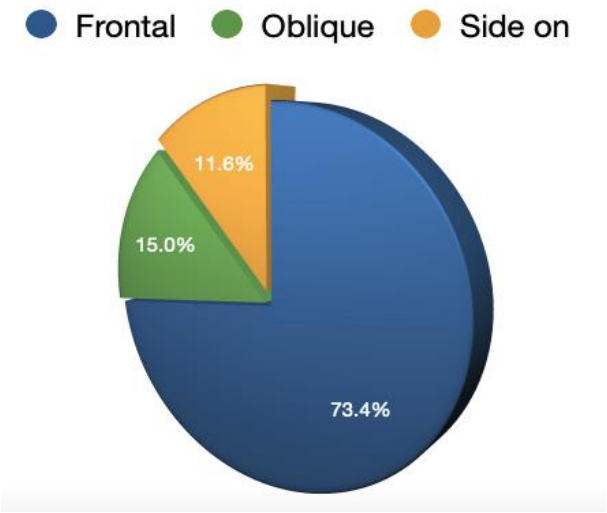


Figure 5: Horizontal angle and involvement

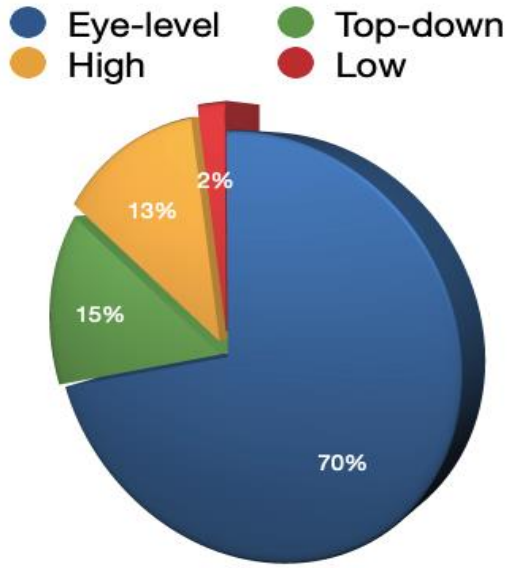


Figure 6: Vertical angle and power

The frontal eye angle has been traditionally associated with the control of eye movements (Pierrot-Deseilligny, Milea and Müri 2004). By taking a frontal point of view, the participants and objects are presented facing the viewer "front on", and the viewer has a maximum sense of involvement with them as part of their own world. Here, it is important to note that while there is little agreement on how to best define the involvement construct, involvement leads to different responses and various degrees of engagement. Involvement influences the amount of mental and physical efforts a viewer puts into the process of communicating with a participant or object presented in the posting. Highly involved viewers will focus more visual attention on the depicted participants or objects, searching for more information of them and processing relevant information in greater detail.

Then, the eye-level angle, according to Kress and van Leeuwen (2006, 140), works more defensive than offensive, designed to elicit no negative feelings from the viewer in social interaction. Taken together, the adoption of both the frontal and eye-level angles alongside that of extreme close-up or close-up in the digital remediation is arguably intended to generate inside the viewer a feeling of being closely connected to a depicted participant or object. Figure 3a is a case in point.

5.2.3. Gaze

Gaze is another technique commonly used in persuasive imagery, because the gaze demands "something from the viewer, demands that the viewer enter into some kind of imaginary relation with" a depicted participant or object (Kress and van Leeuwen 2006, 117–118). When a depicted participant or object looks directly at the viewer, its intention is to seduce them, intimidate them into some course of action, or elicit sympathy from them. By contrast, by no gaze, the viewer is metaphorically positioned as "an invisible onlooker" (Kress and van Leeuwen 2006, 118) of the represented participant or object.

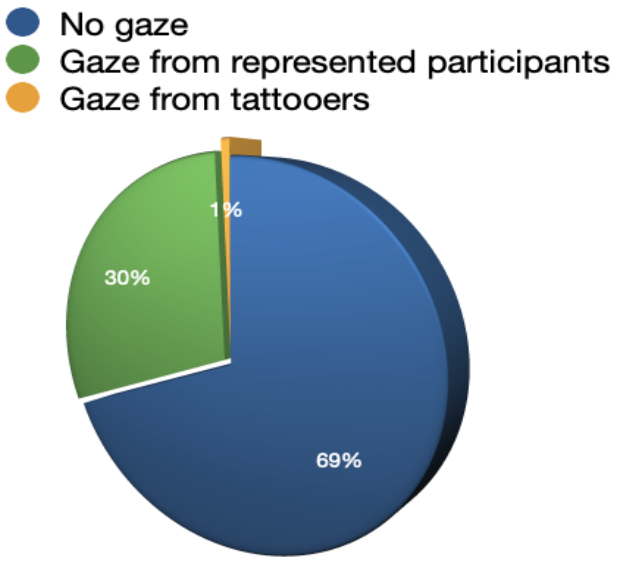


Figure 7: Gaze and image act

As discovered above, the majority of the depicted participants and objects invite the viewer for high involvement and social interaction through frontal and eye-level angles and close shot. Despite this, almost up to 70% of them (see Figure 7) are represented as items of information, objects of contemplation for the viewer's dispassionate scrutiny instead of demanding the viewer to have some kind of imaginary interaction with the depicted. Differently put, among around two thirds of the postings is found no gaze opted to strengthen interpersonal meanings generated by close shot and frontal and eye-level angles of interaction. Nevertheless, by no gaze, it is in congruity with the distinctive peculiarity of blogging as a medium that "offers less of an invitation by the individual to interact with others, but more of a cathartic release or self-presentation what often invites very little to no interaction" (Attrill 2015, 89). The participants or objects were so presented that the viewer is encouraged to feel themselves as an "objective" observer rather than part of the depicted.

This argument developed from the absence of gaze can be further augmented by reinterpreting Kress and van Leeuwen's (2006) interpersonal system of contact. In examining visual narratives of children's picture books, Painter, Martin and Unsworth (2012) maintain that the opposition between the presence and absence of gaze should be construed simply as indicating whether the viewer has been positioned to engage with the depicted participant or object via eye contact, or just to observe the depicted participant or object. Painter, Martin and Unsworth instead proposed the system of "focalization" featured with the primary opposition between contact and observe as the two characteristics for defining the depicted participant or object in visual images. A focalizing choice of "an observing, rather than participating stance" for the majority of the postings, thus, keeps the viewer outside the tattoo world to observe and learn from what goes on within it, which, according to Painter, Martin and Unsworth (2012, 19), in effect gets the viewer preoccupied in the observation, whereby their attitudes towards a depicted participant or object are unconsciously manipulated. Thus said, together with close shot and frontal and eye-level angles, the absence of gaze actually works to force the viewer to look at the depicted participants and objects and accept what is shown without any negotiation. An illustration is presented in Figure 8 where the sign is so closely presented as a visual "offer" (Kress & van Leeuwen 2006) at the eye-level in front of the viewer. According to Painter,

Martin and Unsworth (2012), postings of this kind may be ideologically effective precisely because they do not engage the viewer in overt or convert dialogic negotiation. In other words, viewers are so preoccupied with the depicted participant or object in the posting that their attitudes towards it are unconsciously manipulated.

Figure 8

5.2.4. Modality

In contrast to gaze being optional, modality, an issue of intended truth and reality to be conveyed, is inevitably utilized as an important technique in depicting a participant or object in tattoo form. As mentioned above, the quantitative analysis of modality was conducted in accordance with Painter, Martin and Unsworth's (2012) three styles of character drawing – minimalist, generic, and naturalistic – and the different kinds of alignment. Accordingly, Figure 3a, Figures 3b/3c, and Figure 3d above exemplify the depiction in a naturalistic, generic, and minimalist style respectively. Although not all the postings fit readily into the simple taxonomy, they can still be roughly categorized in terms of these three depiction styles. As shown in Figure 9, 6.8% of the participants and objects in the postings are depicted in the minimalist style, 49.7% in the generic style, and 43.5% in the naturalistic style[37]. In contrast to the minimalist style making the viewer detached from the depicted, the generic style is "likely to more injunctive in nature, implicitly expecting [the viewer] to see themselves in the protagonist role", and the naturalist style invites the viewer to relate and respond to the depicted as "real" (Painter, Martin and Unsworth 201, 33–34). Like Figure 3a, Figure 10 provides a good example of the naturalist style where the viewer is called on to relate and respond to the depicted participants as "real". Both of the figures are also the same in that the depicted participant in them is decontextualized.

[37] With a total number of 59, letters defined as a kind of object were categorized into the naturalist style.

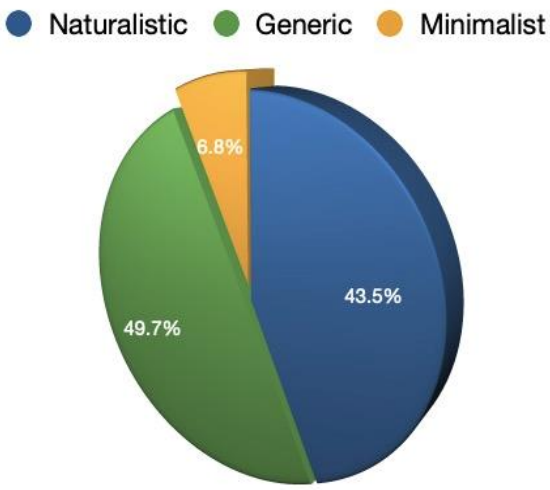


Figure 9: Depiction styles and distribution

Figure 10.

Table 1 presents the quantitative findings regarding the background of the postings in terms of degrees of detail, regardless of the depiction in white-and-grey or color. With the lowest articulation of the background information, more than 52% are completely decontextualized. Among those contextualized, the blurring background amounts to 13%. Taken together, approximately 65% of the depicted participants and objects are themselves most salient and eye-catching, as shown by Figures 3a and 11. In Figure 10, lightness has also been used in the capture of the depicted participant. But only 17.4% of the participants and objects, when captured, are shined upon (see Table 2), in contrast to the relatively predominant practice of background decontextualization.

Background / Tattoo type	black-and-white	blurring	naturalistic	**Total**
Black-and-grey	82	21	65	168
Coloured	71	17	37	125
Total	153	38	102	**293**
Percentage	52.2%	13%	34.8%	**100%**

Table 1: Types of background

Lightness / Tattoo type	added	naturalistic	**Total**
Black-and-grey	32	136	168
Coloured	19	106	125
Total	51	242	**293**
Percentage	17.4%	82.6%	**100%**

Table 2: Lightness of tattoos

To recapitulate, the above analyses of the indicators of photographic techniques suggest that the reviewer is coercively positioned in a specific relationship with the postings, which is mainly accomplished through the predominant application of (extreme) close shot and frontal and eye-level angles in capture to generate the conspicuous depiction of a participant or object. In a posting where its background is naturalistically contextualized, the possibility remains for the viewer to foster the same feeling of being closely connected to the decontextualized postings due to being closely and directly positioned front. This way of visual positioning is powerful in that the viewer is obliged to at least temporarily accept in order to process the depicted participants and objects. Secondly, at first sight, not all the photographic techniques discussed above necessarily work in the direction for the same interpersonal meanings. For example, the positioning of the viewer as the observer of a depicted participant or object due to the absence of gaze is clearly divergent from the deep involvement of the viewer with a depicted participant or object created by (extreme) close shot and frontal and eye-level angles. So is the use of decontextualization versus that of frontal angel and close shot. The visual divergences which keep the viewer slightly off balance appear to act as a metaphor for a somewhat ambiguous viewing position created by interpersonal choices, which paradoxically either introduce ready alignment with a depicted participant or object or let it alone. Nevertheless, in whichever cases the adoption of frontal and eye-level angles and close shot, as analyzed above, introduces the limits to negotiability on the viewer's side.

It is important to accentuate that the foregoing discussion is not meant to declare the unlikelihood for sociocultural meanings of tattooing to be negotiated. Instead, the negotiability of sociocultural meanings of tattooing is visually maximized by virtue of the fact that tattoos have been digitally remediated in such ways that the identification of any initial common ground itself now becomes a substantial challenge. This leads to the next point worthy of an emphasis; that is, the photographic techniques sometimes are inextricably intertwined with each other. A good case in point is one of the above-mentioned examples of visual divergence, where both frontal angle and (extreme) close shot are the prerequisites for encouraging ready alignment with a depicted participant or object in the contextualized postings. As such, it is cogent to argue that the working of the postings in negotiating sociocultural meanings of tattooing in China is less indebted to decontextualization than

the exploitation of frontal angle and (extreme) close shot that impose very specific constraints on the kinds of information that can be represented and processed and the subsequent mediation of affect.

6. Conclusion

Contextualized within the hegemonic order where long-standing stigmatized associations of tattoos are still persistent in China, this article has explored how photographic images of corporeally mediated tattoos alongside the blog can be drawn upon together for personal expression and negotiating sociocultural meanings of tattooing. In alignment with the recent theorization of semiotic remediation in what has been called the post-digital era by underlining the pivotal role of digital remediation in delivering political messages, the article concentrated primarily on photographic techniques harnessed and exploited in capturing corporeally mediated tattoos for a quantitative investigation into semiotic means of digital remediation that make it possible self-expression and negotiating sociocultural meanings of tattooing at the blog. While this study didn't offer a nuanced account of how specific affective formations take place via specific practices of photographic techniques, its quantitative findings support effects of photo manipulation on identity presentation, meaning negotiation, and the viewer's attitudes.

As analyzed above, after being tactfully remediated with the selected photographic techniques and displayed at the blog, corporeally mediated tattoos per se are endowed with new meaning potentials to negotiate or counter certain socialized prejudice and ideologies in addition to self-expression. Not only does this finding provide empirical support for digital remediation as repurposing; it would also be awfully inspiring to socially stigmatized and marginalized (sub)groups who want to circumvent social, cultural and political barriers to communicate and make their stories known to more people. This study, thus, is added to existing, mostly qualitative, tattoo studies that generally draw on interviews with tattoo wearers, enthusiasts and artists for an account of tattoo narratives in connection to personal expression.

This study is premised both on the notion of visual manipulation achievable via photographic techniques harnessed and exploited in the digital remediation processes and on Prior and Hengst's (2010) articulation of "semiotic remediation as repurposing". As a conclusion, it may also be crucial to point out that digital remediation that purportedly eludes conscious perceptions might be impervious to visually literate viewers (cf. Lazard et al. 2018; Lazard, Bock and Mackert 2020). Moreover, tattoo narratives in postings are most profitably not studied as decontextualized, denotational texts, since they are dynamically reconfigured by the interaction between the viewer and the posting. In order to comprehensively appreciate digital remediation and its association with visual manipulation, the real perceptions of the viewer or their spontaneous reactions to the postings should be considered too. Thus, it is desirable to conduct a close interview and examination of how the viewer perceives and makes sense of the postings, since their perceptions also depend on the embodied positionality, as well as engage with the power implications of affect in visual manipulation. Finally, how blogs are mobilized by the social (sub)group of Chinese tattoo wearers and enthusiasts to distinguish themselves from others and gradually gain social recognition in China is also worthy of study. This is just because after being posted at the blog, digitally remediated tattoos likely develop a virtual community of practice (Wenger, 1998) where a collective identity is presented and solidified.

References

Ahmed, S. 2004. *The cultural politics of emotion*. New York: Routledge.

Alvarez, P. 2020. Indigenous (re)inscription: Transmission of cultural knowledge(s) through tattoos as resistance. In S. T. Kloβ (ed.), *Tattoo Histories.* London: Routledge, 157–175.

Androutsopoulos, J. 2021. Polymedia in interaction. *Pragmatics and Society* 12 (5): 707–724.

Ang, T. 2019. The big cover-up: Tattoo culture in China. https://english.ckgsb.edu.cn/knowledges/tattoo-culture-in-china/, accessed 17 November 2021.

Atkinson, M. 2002. Pretty in ink: conformity, resistance, and negotiation of women's tattooing. *Sex Role* 47 (5/6): 219–235.

Attrill, A. 2015. *The Manipulation of Online Self-presentation: Create, edit, re-edit and present.* New York: Palgrave Macmillan.

Bateman, J. A. 2021. What are digital media? *Discourse, Context & Media* 41, 100502. https://doi.org/10.1016/j.dcm.2021.100502.

Bengtsson, A., Ostberg, J. & Kjeldgaard, D. 2005. Prisoners in paradise: Subcultural resistance to the marketization of tattooing. *Consumption, Markets and Culture* 8 (3): 261–274.

Bolter, J. D. & Grusin, R. 2000. *Remediation: Understanding new media*. Cambridge, UK: MIT Press.

Bucholtz, M. 2003. Sociolinguistic nostalgia and the authentication of identity. *Journal of Sociolinguistics* 7 (3): 398–416.

Couldry, N. 2012. Media as practice. In N. Couldry (ed.), *Media, Society, World: Social theory and digital media practice*. Cambridge, UK: Polity, 33–58.

Davey, G. & Zhao, X. 2019. Tattoos, modernization, and the nation-state: Dai Lue bodies as parchments for symbolic narratives of the self and Chinese society. *The Asia Pacific Journal of Anthropology* 20 (2): 165–183.

DeMello, M. 2000. *Bodies of Inscription: A cultural history of the modern tattoo community.* Durham, UK: Duke University Press.

Deumert, A. 2014. *Sociolinguistics and Mobile Communication*. London: Routledge.

Djonov, E., Tseng, C. & Lim, V. L. 2021. Children's experiences with a transmedia narrative: Insights for promoting critical multimodal literacy in the digital age. *Discourse, Context & Media* 43: 100493. http://doi.org/10.1016/j.dcm.2021.100493.

Emme, M. J. & Kirova, A. 2005. Photoshop semiotics: Research in the age of digital manipulation. *Visual Arts Research* 31: 145–153.

Fisher, J. A. 2002. Tattooing the body, marking culture. *Body & Society* 8 (4): 91–107.

Hiramoto, M. 2015. Inked nostalgia: displaying identity through tattoos as Hawaii local practice. *Journal of Multilingualism and Multicultural Development* 36 (2): 107–123. http://doi.org/080/01434632.2013.804829.

Ifukor, P. 2011. Linguistic marketing in "…a marketplace of idea": Language choice and intertextuality in Nigerian virtual community. *Pragmatics and Society* 2 (1): 110–147. http://doi.org/10.1075/ps.2.1.06ifu.

Ivković, D. & Lotherington, H. 2009. Multilingualism in cyberspace: Conceptualizing the virtual linguistic landscape. *International Journal of Multilingualism* 6 (1): 17–36. https://doi.org/10.1080/14790710802582436.

Jaworska, S. 2018. 'Bad' mums tell the 'untellable': Narrative practices and agency in online stories about postnatal depression on Mumsnet. *Discourse, Context and Media* 25: 25–33.

Jaworski, A. & Thurlow, C. 2010. Introducting semiotic landscapes. In A. Jaworski and C. Thurlow (eds.), *Semiotic Landscapes*. London: Continuum, 1–40.

Kallen, J., Dohnnacha, E. & Wade, K. 2020. Online linguistic landscapes: Discourse, globalization, and enregisterment. In D. Malinowski & S. Tufi (eds.), *Reterritorializing linguistic landscapes: Questioning boundaries and opening spaces*. London: Bloomsbury, 96–116.

Koller, V. & Bullo, S. 2019. "Fight like a girl": Tattoos as identity constructions for women living with illness. *Multimodal Communication* 8 (1), 20180006. http://doi.org/10.1515/mc-2018-0006.

Kress, G. & van Leeuwen, T. 2006. *Reading Images: The grammar of visual design* 2nd ed. London: Routledge.

Lane, D. 2014. That's all folks: An analysis of tattoo literature. *Sociology Compass* 8 (4), 398–410.

Lazard, A. J., Mackert, M.S., Bock, M. A., Love, B., Dudo, A. & Atkinson, L. 2018. Visual assertions: effects of photo manipulation and dual processing for food advertisements. *Visual Communication Quarterly* 25 (1), 16–30. https://doi.org/10.1080/15551393.2017.1417047

Lazard, A. J., Bock, M. A. & Mackert, M. S. 2020. Impact of photo manipulation and visual literacy on consumers' responses to persuasive communication. *Journal of Visual Literacy* 39 (2): 90–110.

Van Leeuwen, T. 2005. Introducing Social Semiotics. London: Routledge.

Lei, D. P. 2009. The blood-stained text in translation: tattooing, bodily writing, and performance of Chinese

virtue. *Anthropological Quarterly* 82 (1): 99–128.
Limatius, H. 2019. “I’m a fat bird and I just don’t care”: a corpus-based analysis of body descriptors in plus-size fashion blogs. *Discourse, Context & Media* 31, 100316. https://doi.org/10.1016/j.dcm2019.100316.
Limatius, H. 2020. “I think she’s truly beautiful”: celebrity, gender and body positivity in plus-size fashion blogs. *Participations: Journal of Audience & Reception Studies* 17 (2): 372–392.
Mallya, D., & Susanti, R. 2021. Theorizing race, marginalization, and language in the digital media. *Communication & Society* 34 (2): 403–415.
Martin, C. W. 2019. *The Social Semiotics of Tattoos: Skin and self.* London: Bloomsbury.
McLuhan, M. 1964. *Understanding Media: The extensions of man.* New York: McGraw-Hill.
Messaris, P. 1994. *Visual “Literacy”: Image, mind, and reality. Boulder*, CO: Westview Press.
Messaris, P. 2012. Visual “literacy” in the digital age. *The Review of Communication* 12 (2), 101–117.
Miller, C. R., & Shepherd, D. 2004. Blogging as social action: a genre analysis of the weblog. *Into the Blogosphere Articles*. University of Minnesota. https://conservancy.umn.edu/handle/11299/172818
Myers, G. 2010. *Discourse of Blogs and Wikis*. London: Continuum.
Page, R. 2011. *New Narratives: Stories and storytelling in the Digital Age*. Lincoln: University of Nebraska Press.
Page, R. 2018. *Narratives Online: Shared stories in social media*. Cambridge, MA: Cambridge University Press.
Painter, C., Martin, J. R. & Unsworth, L. 2012. *Reading Visual Narratives: Image analysis in children’s picture books.* Sheffield, UK: Equinox.
Patterson, M. 2018. Tattoo: marketplace icon. *Consumption Markets & Culture* 21 (6): 582–589.
Peck, A. & Stroud, C. 2015. Skinscapes. *Linguistic Landscape* 1 (1): 133–151.
Peck, A. & Williams, Q. 2018. Skinscapes and frictions: an analysis of Zef Hip-Hop ‘Stoeka-style’ tattoos. In A. Peck, Q. Williams & C. Stroud (eds.), *Making Sense of People and Place in Linguistic Landscapes*. London: Bloomsbury, 91–106.
Pierrot-Deseilligny, C., Milea, D. & Müri, R. 2004. Eye movement control by the cerebral cortex. *Current Opinion in Neurology* 17, 17–25.
Prior, P. A. & Hengst, J. A. (eds.). 2010. *Exploring Semiotic Remediation as Discourse Practice*. New York: Palgrave Macmillan.
Rosenblatt, D. 1997. The Antisocial skin: Structure, resistance, and “modern primitive” adornment in the United States. *Critical Anthropology* 12 (3), 287–334.
Roux, S., Peck, A. & Banda, F. 2019. Playful female skinscapes: body narrations of multilingual tattoos. *International Journal of Multilingualism* 16 (1): 25–41.
Schildkrout, E. 2004. Inscribing the body. *Annual Review of Anthropology* 33: 319–344.
Scollon, R. & Scollon, W. S. 2003. *Discourses in Place: Language in the material world.* New York: Routledge.
Shouse, E. 2005. Feeling, emotion, affect. *M/C Journal* 8 (6). https://doi.org/10.5204/mcj.2443
Tagg, C., & Lyons, A. 2021. Polymedia repertoires of networked individuals: A day-in-the-life approach. *Pragmatics and Society* 12 (5), 725–755.
Trainer, S., et al. 2016. The fat self in virtual communities: success and failure in weight-loss blogging. *Current* Anthropology 57 (4), 523–528.
Walzer, A. & Sanjurjo, P. 2016. Media and contemporary tattoo. *Communication & Society* 29 (1), 69–81.
Wee, L. 2016. Situating affect in linguistic landscape. *Linguistic Landscape* 2 (2), 105–126.
Wenger, E. 1998. *Communities of Practice: Learning, meaning, and identity.* New York: Cambridge University Press.

A Methodological Guide to Building Digital Materials for the Sociophonetic Research of Vowels

Simon Gonzalez
The Australian National University
E-mail: u1037706@anu.edu.au

Abstract

The current stage of sociophonetic research entails a specialised combination of skills. This requires researchers to have a solid understanding of phonetic phenomena as well as developed computational capabilities which are used for the analysis of speech data. Specific skills are applied at every stage of the research process: data collection, preparation (wrangling), visualisation, and analysis. Workflows in the field vary depending on the available tools and the established skills of the people involved. In this paper, we present a selection of data analysis tools used in the creation of an online web app for presenting digital materials used in sociophonetic research. In this sense, the work is more methodologically oriented. The tools are arranged in such a way that it proposes a workflow that captures the relevant tasks present in most of the projects of similar nature. This is also intended to be a contribution to streamlining the creation of digital resources that can be used in any sociophonetic research project. The tools are developed within the R environment and the deployment of materials is done through Shiny Apps, within RStudio. For the speech processing tasks, we also make use of the Praat software, which combined with R packages, can deal with different tasks: data preparation, speech processing, measurements, visualisation, and app deployment. The final product is an open-source web application. The final product can be freely accessed and applied in any study in which they can be adapted.

Keywords: digital materials, data analysis tools, sociophonetics, web apps, R, Shiny

1. Background

Innovation is at the front line of any research field in the current age. The humanities are not the exception. Digital technologies have influenced how we carry out research across the humanities, which is now a field on its own: *Digital Humanities*. It has been defined as the intersection between traditional humanities and computational methods (Burdick et al. 2012). This has strong implications, especially with the "computational methods" aspect of the definition, which are defined as "mathematical models used to numerically study the behaviour of complex systems by means of a computer simulation." (Nature Portfolio 2021). This definition accurately encapsulates the reality for speech research within the Humanities. Language is a complex system, and the most streamlined way in which it is studied is through the digitisation of the speech signal. Once digitised, it can be analysed in an array of ways, especially in the revolution of speech processing leveraged by Automatic Speech Recognition (ASR) technologies. This has facilitated both the creation and access of speech corpora through the internet.

1.1. Speech Corpora

Speech corpora can be accessed through several resources freely available which have been put together by different research projects. Two representative examples can be found in Butryna et al. (2019) and Mazumderm et al. (2021). A seminal corpus which has become a landmark in phonetic research is the *Speech Accent Archive* (Weinberger & Kunath 2015), which gathers speakers from around the world recording themselves reading the same prompt and it is shown below:

> *Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob.*

We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

This paragraph is phonetically rich since it contains all the sounds of English. A great advantage of this resource is that it offers varied phonetic contexts, which can be maximised for linguistic analysis. For instance, similar phonetic features can be compared across a range of contexts such as word positions and preceding and following segments Another characteristic is that the database gives demographic information for each of the speakers. These include age, gender, place of birth, and native language. This resource is now widely used in technological and engineering studies (c.f. Rajan et al. 2022; Ahamad et al. 2020; Ahmed et al. 2019).

The *Archive* is open source, and it can be accessed through the internet. There are different ways in which the dataset can be downloaded for local use and analysis[38]. The first option is to manually download the audio files. However, downloading a large number of files available can be a time-consuming process, even if a subset is selected. The second option is to do it through a computational approach, where the files can be downloaded and stored using an R (R Core Team 2021) or Python (Van Rossum & Drake 1995) script, which makes it easier to automate the downloading process. This is developed more in Section 2.1. (Data Harvesting and Preparation). Once downloaded, WAV files require heavy processing for obtaining phonetic results. In current analysis workflows, there is a combination of tools and techniques that allow researchers to extract acoustic information and examine the data.

1.2. Stages in the Sociophonetic Research Process

The process of analysing speech data in sociophonetic research is not the same across different projects. First, the available data (or lack thereof), is a crucial part of the process. As is the case of most of the world's languages, the starting point is as the data collection stage. In the current work, we start at a point where the data is already publicly available. Second, the required tools and the stages defined vary depending on the theoretical framework and the perspectives of the researchers involved. Here we describe a four-staged process that aims to capture the main steps in speech analysis based on our experience and based on many other studies in the field. It is not intended to be exhaustive or authoritative. Another important observation is that these stages are not always contiguous, since there are researchers who prefer to carry out certain tasks in a more overlapping way, even having cycles in some cases, i.e., coming back to a previous stage and do the process again. We also try to capture the common strategies regardless of the methodological approach, such as visualisations and the pre-processing of the data, as shown in Figure 1.

Figure 1. Main Stages in the Sociophonetic Research process.

[38] The database can also be accessed through a Kaggle project in https://www.kaggle.com/datasets/rtatman/speech-accent-archive

1.2.1. Data Preparation

In this stage, speech features are extracted from audio recordings, generally from WAV, MP3[39], or any other audio format. The audio files are generally accompanied by their corresponding transcriptions, and they can be in a multiple of formats that allow time-stamped information of the transcribed content, including text files (TXT), tab separated values (TSV), comma separated values (CSV), *Elan* Files (EAF), and *Praat* files (Boersma 2022) (TextGrid). The key component here is to have a time alignment with the corresponding audio file, as shown in Figure 2. The alignment can be at different levels, including sentence or utterance level, but it can be as granular as at the phonemic level.

In the case of sociophonetic research, a great number of studies focuses on phonemic segmentations done by force-aligning the data, which outputs the transcriptions with segmentations at the phonemic level for both consonants and vowels. Once forced-aligned, the acoustic features are extracted and are saved as numeric values. Two relevant acoustic measurements are duration (time) and formant information (Hz), which are then analysed looking at sociolinguistic factors such as gender, age, socioeconomic status, and other relevant factors. Other acoustic measurements are also important for phonetic analysis, including pitch, intensity, Mel-Frequency cepstrum (MFCC), and others. However, we chose to present analysis on duration and formant information based on practical terms due to the limitations of the scope of the application. Also, duration and formant information are robust features widely used in phonetic analysis that can reliably describe acoustic patterns in the data.

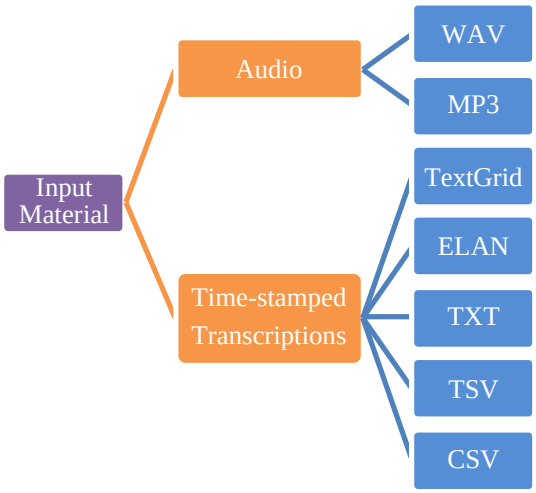


Figure 2. Main File Formats in the Data Processing.

1.2.2. Data Visualisation

Within sociophonetic research, data visualisation can take any path as required by the study. However, five types of commonly used visualisations are bar plots, box plots, scatter plots, line plots, and Cartesian vowel spaces (See Figure 3). Bar plots and scatter plots are generally used for displaying numeric differences between categorical groups, e.g., consonant duration differences between speakers from different regions. Scatter plots are generally used to visualise correlations between two

[39] In sociophonetic research, there is a preference for avoiding, if possible, using MP3 files if a WAV file is available. The reason is that the acoustic quality tends to be compromised due to the compression of the file. This can affect sound features, such as the frication of sibilant sounds. However, it is not uncommon to find studies done on MP3 files, as long as the files are treated properly, and the acoustic features are not compromised. Caution is therefore in order when applying any kind of acoustic analysis on audio files that were originally recorded as MP3.

numeric measurements, e.g., visualise vowel rising across time, which is generally used to measure sound change, where we measure vowel rising/lowering/fronting/backing in different generations of speakers. Line plots are used in phonetic plots to visualise dynamic formant values for vowels or sonorant consonants. Finally, Cartesian vowel spaces allow the visualisation of vowels, both as static points or as dynamic trajectories. This is by far the most widely used visualisation type in vowel analysis.

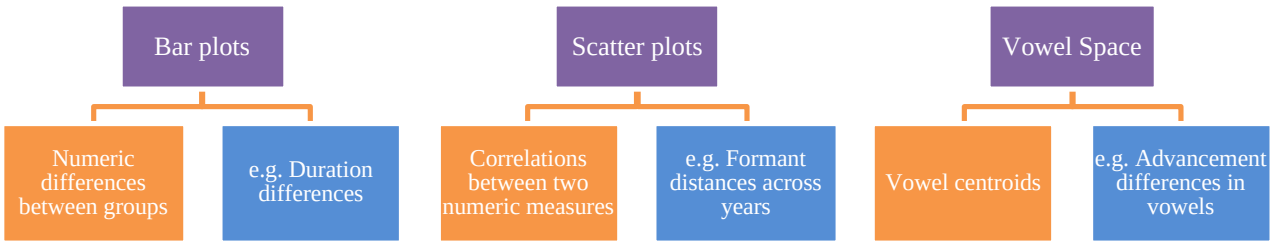


Figure 3. Visualisation Types and Applications in the Workflow.

1.2.3. Data Analysis

The data analysis approaches also vary depending on the main purpose of a given study. However, the following types of analyses here presented are common practice in the field of sociophonetics[40] (c.f. Diskin et al. 2019; Docherty & Foulkes 1999; Docherty et al. 2015), and we limit the analysis in this current work to the analysis of vowels in English. We base our approaches on Harrington (2010). The types of analyses below are not presented in a hierarchical order, but rather the order was chosen for practical reasons. The first one is the analysis of formants and formant trajectories, which are used to estimate acoustic activity in the vocal tract. This is based on Hz values and allow doing comparisons of speakers across different sociophonetic groups. We use formants to measure differences in terms of the vertical axis (vowel rising, for example, rising of low vowels to central vowel positions), or the horizontal axis (vowel fronting, for example, back vowels move to more fronted positions) (c.f. Diskin et al. 2019; Docherty et al. 2019).

The second type of analysis is vowel ellipses. This type of analysis is based on the premise that even when there are intra- and inter-speaker differences in the productions of vowels, realisation of vowels revolves around an articulatory target that can be captured in "areas" of production and not necessarily "points" of articulation. In this sense, vowel articulations in a language revolve around centroids that are relative to each other. These ellipses show confidence regions for each of the vowels. Confidence regions are a two-dimensional generalisation of a confidence interval and are represented as ellipses, which are placed around centroids – the point of central tendency of a distribution (Amin et al. 2014). Vowel ellipses are relevant acoustic representations because a confidence region is useful for identifying the location and spread of a vowel category, and this can be used to compare across vowel categories, and phonological contexts. This is crucially relevant when analysing vowel mergers – when different vowels merge into one (c.f. Labov et al. 1991; Maclagan & Gordon 1996; Gordon & Maclagan 2001; Hay et al. 2006; Diskin et al. 2019), and vowel splits – when one vowel splits into two (c.f. Docherty & Foulkes 1999; Hughes et al. 2005; Piercy 2011; Zellou & Scarborough 2019; Jansen & Braber 2020).

Vowels are analysed in both static and dynamic ways. For static analyses, one point of reference within each segment is set at given proportions of the duration of the segment (Watson & Harrington

[40] In terms of analysis, the field of Sociophonetics shares approaches from Speech Corpora analysis.

1999; Van Heuven et al. 2002; Cox 2006). For dynamic analyses, the aim is to capture the movement of vowels across time (McDougall 2006; Haddican et al. 2013; Docherty et al. 2015). Among the methods used are Euclidean Distances, which measure the distances between two points in the vowel, generally the points identified as the onset and the offset[41]. In our case, we define these points at the 20% and 80% for diphthongs and 50% for monophthongs, following Van Heuven et al. (2002) where specific points are defined. These analyses examine vowels in their phonological context, generally looking at Manner and Place of Articulation, Voicing, and Type (vowel or consonant) of previous and following segments.

1.2.4. Visualising Data and Presenting Results

Materials in sociophonetic research are generally presented within the framework of *Speech Corpora*, which aims to build corpora and provide users with interfaces that allow access to the data. The most crucial characteristic is to present a system that facilitates the searching and listening to given segments or portions within individual recordings. This is crucial because it lets users access the raw data from where the analysis and visualisations come from. Other options include the visualisation of features beyond the metadata as such, for example, visualising sound durations, vowel formants, and vocalic spaces. Analysis results can also be presented, choosing from an array of options such as tables (fixed and interactive), summary plots, and acoustic model summaries.

In terms of the platform and programming language used, both R and Python are widely used for these purposes. In our experience, R is strongly used in sociophonetic research, and it also offers the development and deployment of online apps through Shiny apps (Chang et al. 2019). These apps are relatively straight forward to develop, and the results give users great control over applications that are both efficient and interactive. Additionally, Shiny applications are intrinsically reactive, which is invaluable when interacting with online apps.

1.2. Main Goal of the Paper

The main goal of this paper is to develop a methodological framework whose end is the visualisation of speech materials and their analyses within an online application. The intended audience is language researchers working on vocalic analysis, and computational linguists who can extend the current code to include more analytical power, customising it to new research needs. The focus of our analysis in on English vowels. We present available tools, which are used to gather online available data and build digital materials to be shared through digital means. The contribution to digital humanities is two-fold. First, the computational tools developed here equip users with strategies to make effective use of digital materials available for their research. This can be applied to written texts, audio files, images, and any other relevant format. The second intended contribution is to present the available options for building online applications which can present both raw data and processed results from different analyses. These are tools that can be accessed freely due to their open-source nature, which can boost the way we deal with online materials available in the current digital age.

2. Building the Tools

In this section, we go through the different tools and R packages that are used to process the data. We

[41] Many studies define target points in the trajectory. In this current work, we base our points of pre-specified and fixed points in the trajectory.

follow these stages in the overall process: data collection, data processing, data visualisation, data analysis and deployment.

2.1. Data Harvesting and Preparation

Online available data can be harvested from RStudio (RStudio Team 2021). The package widely used for this purpose is *rvest* (Wickham 2021). This package allows accessing and downloading available data from web pages. It does this by reading the URL content in HTML, then its content can be accessed through pipelines in the R code. In our case, we started reading from the *Speech Accent Archive* page where the languages were indexed (https://accent.gmu.edu/browse_language.php) as the root. This had a total of 389 languages in the database. Using the *rvest* package, we downloaded the audio files with their demographic information: native language, gender, region, and country (See Figure 4 below).

Figure 4. Source website from which the data is obtained.

Since the original reading in the audio files came from a fixed prompt, we then proceeded to create a corresponding TextGrid using the prompt as the main text and the time duration for each audio file. We did this using the *rPraat* R package (Bořil & Skarnitzl 2016) which allows the creating of Praat TextGrids within R. One of the uses of having audio files with their corresponding transcriptions in TextGrids is that they are then used as inputs for the force-alignment process. In this study, we used the Montreal Forced-Aligner (McAuliffe et al. 2017). The output were TextGrids forced-aligned at the phonemic level, as shown in Figure 5 below[42].

[42] The forced-alignment quality depends on may factors, such as audio quality, transcription quality, language data available, and amount of data available, among others.

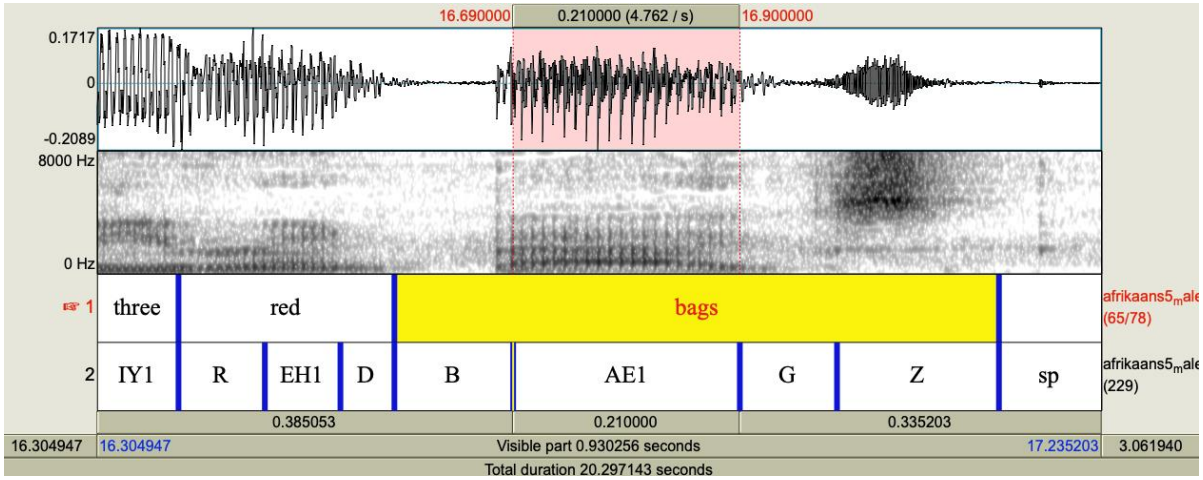


Figure 5. Sample output from the forced-alignment process. Each segment is time-aligned to the audio file.

2.2. Acoustic Feature Extraction

Once the data was forced-aligned, we proceeded with the speech processing stage. For this, we used a combination of R scripts and Praat Scripts. There is an array of useful and powerful R packages that can be used for linguistic and sociophonetic analysis (See Table 1 for R packages used), e.g *emuR* (Winkelmann et al. 2021), *PhonTools* (Barreda 2015), *PraatR* (Albin 2014). However, here we present an approach that allows users to have the most customised control on the analysis. It is important to note that this process can also be done using only Praat Scripting (or Python or any other programming language of choice), but since our main framework was within R, we decided to do the data preparation in R, so it can be better connected with further stages in the workflow. Within R, we used the functionality from the *dplyr* package (Wickham et al. 2022), which has a wide range of functions to wrangle any data within R. This package is part of *tidyverse* (Wickham et al. 2019), and in our experience, this is by far one of the most powerful packages (or set of packages) for data wrangling and visualisation in R. We prepared the data in such a way that we had all information available for each of the phonemes for every speaker: segment (vowel or consonant), manner and place of articulation, and voicing of previous and following segments, including spaces. We also added the duration information (duration, onset, offset, midpoint) across 11 equidistant points for each segment. This was done to capture dynamic information of the vowels. We then created other column variables which we labelled but left empty: Formant Information (F1 and F2), Pitch and Intensity. These empty columns were then filled by importing the audio files into a Praat Script. This script first created the corresponding formant, pitch, and intensity files for each of the WAV files. These were then used to extract the corresponding acoustic values at the specified time points in the previous wrangling stage within R. The output from this script was a dataset with the filled corresponding acoustic values for each of the 11 points in each segment.

Package	Usage	Authors	URL
dplyr	*Data wrangling*	*Wickham, H., François, R., Henry, L. & Müller, K*	*https://dplyr.tidyverse.org*
emuR	*Acoustic processing*	*Winkelmann, R., Jaensch, K., Cassidy, S. & Harrington, J.*	*https://github.com/IPS-LMU/emuR*
PhonTools	*Phonetic/Phonological processing*	*Barreda, Santiago*	*https://github.com/santiagobarreda/phonTools*
vowels	*Formant processing*	*Kendall, T. & Thomas, E.-R.*	*http://blogs.uoregon.edu/vowels/*
ggplot2	*Visualisations*	*Wickham, H.*	*https://ggplot2.tidyverse.org*
plotly	*Interactive plots*	*Sievert, C*	*https://plotly-r.com*
shiny	*App development*	*Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J*	*https://CRAN.R-project.org/package=shiny*

Table 1. Main R packages, their corresponding functions used, and authors.

2.3. Speech Analysis

The Praat script completed the file that was the main core of analysis for the creation of this digital material. In this file, we had stored both the demographics information and the acoustic data. We then proceeded to analyse the vowels following standard processes in the sociophonetics field. For the normalisation of formant trajectories, we used the *vowels* package (Kendall & Thomas 2018). We used the Lobanov normalisation method (Lobanov 1971) from the *vowels* package[43]. The normalised values were then used to measure vocalic duration differences and formant comparisons between groups (e.g. native language, country, region, city). Monophthongs are compared at the midpoint (50%) and diphthongs at 20% and 80% of their full trajectories. An important consideration here is that these measurements are done interactively whenever the app users change the comparisons groups (an example is shown in Figure 6). In this way, we make strong use of the interactivity functionality of Shiny apps. This is an effective way of combining these tools. Instead of creating multiple individual files with mean differences or comparisons across groups beforehand, we do this based on the requests from the online users in an interactive way. The way results are displayed is explained in the following section.

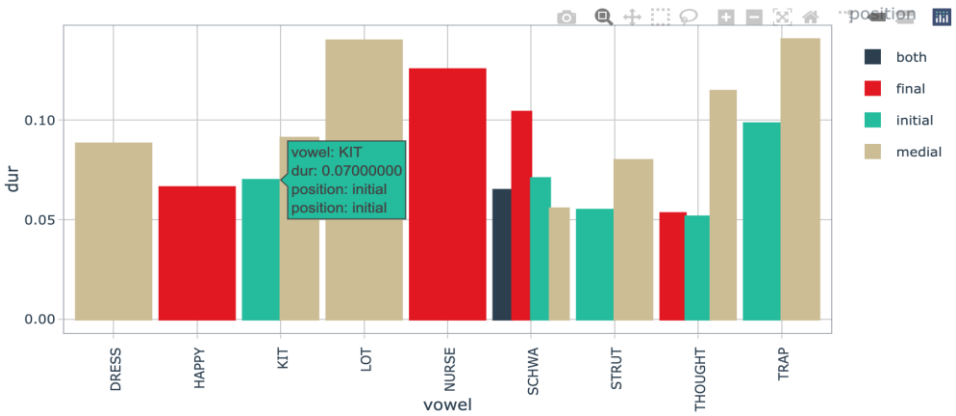


Figure 6. Durational differences across vowels based on their phonological position. More detailed information on durational differences, such as contexts, are addressed in the app.

[43] There is a wide range of other methods for vowel normalisation, including Bark, Labov, Nearey, Watt and Fabricius.

2.4. Data Visualisation

R offers an incredible array of libraries that are used for visualisation purposes. There has also been a sharp increase of libraries that help interactivity and that are used in Shiny apps[44]. Among these libraries, *ggplot2* (Wickham 2016) stands out from the others due to its legacy, vast documentation, highly customisable power, and, very importantly, great aesthetics. We therefore chose *ggplot2* because it can be adapted to visualise the high complexity of the speech data we represent in the digital materials. *ggplot2* also has many extensions[45] that are frequently expanded and are highly flexible in their implementation. One limitation that tends to be reported in relation to *ggplot2* is its adaptability for interactive visualisations. There are two packages that are commonly used to add interactivity to *ggplot2*: *ggiraph* (Gohel et al. 2021) and *plotly* (Sievert 2020). Choosing one or the other depends on the preference of users, both with advantages and disadvantages. In our case, we use *plotly* to make *ggplot2* interactive in our Shiny app. It is our personal opinion that it is a more straightforward process than using *ggiraph*, since we use the `plotly::ggplotly()` function to make *ggplot2* fully interactive. It is important to note that there are some limitations, given that there are plot layers in *ggplot2* which have not been implemented in *plotly* yet, e.g. the `geom_label()` layer.

In terms of the visualisation types, we use bar plots to visualise vocalic duration differences, and Cartesian vocalic spaces to visualise vowel comparisons, both monophthongs and diphthongs, across groups selected by the app users. We followed the conventional plot of F2 values on the x-axis (reversed) and F1 values (reversed) on the y-axis, which correlate with the activity in the vocal tract (See Figure 7). This allows users to compare the results with other studies following the same convention. In addition, all the wrangled data used for the analysis and visualisation can be downloaded as a CSV file. The main purpose is to facilitate researchers accessing data that can be used in analyses outside the app.

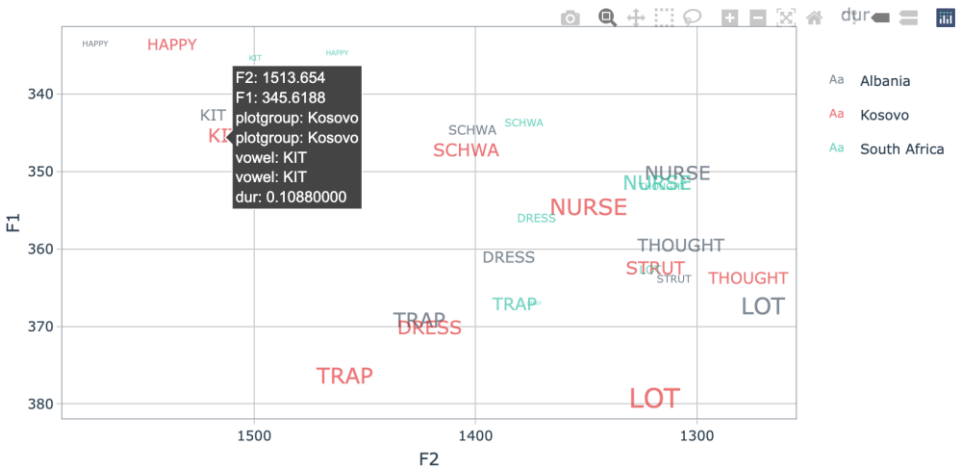


Figure 7. Visualisation of vowel space differences across different countries of origin. The vowel locations represent the mean formant information at temporal mid-points across speakers. The size represents the mean durations across speakers of the same country.

[44] For a collection of relevant packages used for visualisation within Shiny apps, please visit: https://www.htmlwidgets.org/
[45] For a collection of current ggplot2 extensions, please visit: https://exts.ggplot2.tidyverse.org/

3. Features, Structure, and Organization of the User Interface

This project aims to present the materials in a holistic way. In this sense, the goal is to give the user a full experience with the data, where they listen and see the linguistic features visualised through a user-friendly interface. This is done to maximise the use of current technologies, where we aim to break the disconnection between the source data and what it is telling through powerful visualisation tools. It also allows users to combine analysis with visualisations in a more effective way, where the two processes are not separated but are part of the same analytical continuum.

On the left side, the user can select what to listen to by choosing the native language, country, region, city, and gender of the speaker(s). After making the selection, the main section in the right panel updates all the fields: playing the audio file(s) and the visualisation fields. The app also has the capability of downloading the audio files in WAV format, the graphics in JPG format, and the numeric data of the selection in a CSV format (See Figure 8). These can be used for analysis, and the plots can be used for sharing and publication purposes. Here, we aim to give users the power to maximise the data in the app without needing any knowledge on programming or data processing. In this way, users can focus on what drives their interest and being more effective in their exploration process.

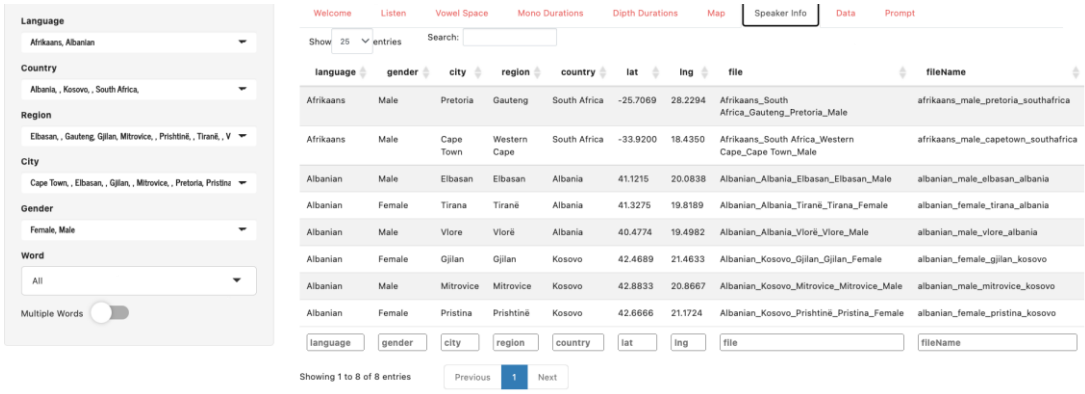


Figure 8. Display of table showing demographics information of selected files.

4. App Deployment

The material was deployed using the Shiny web apps infrastructure (See Figure 9). The code can be accessed in this GitHub repository: https://github.com/simongonzalez/WorldEnglishApp. It publishes Java-based apps from RStudio. Shiny apps require a subscription, with five options to choose from. They range from a FREE option to a PROFESSIONAL subscription[46]. We used the BASIC Subscription since it has a performance boost feature. Compared to the less advanced subscriptions, it offers multiple worker processes per application, and it gives more RAM, which is crucial given the amount of data we process interactively. However, if the app is not envisioned to be used by many people at the same time, or the data processing is not large, then the FREE and STARTER subscriptions are other options. The final product is then a fully functional web app that can be accessed anywhere around the globe, by phoneticians, sociophoneticians, language teachers, or any user interested in observing patterns of accented English around the world.

[46] For a full description of all the options, visit: https://www.shinyapps.io/

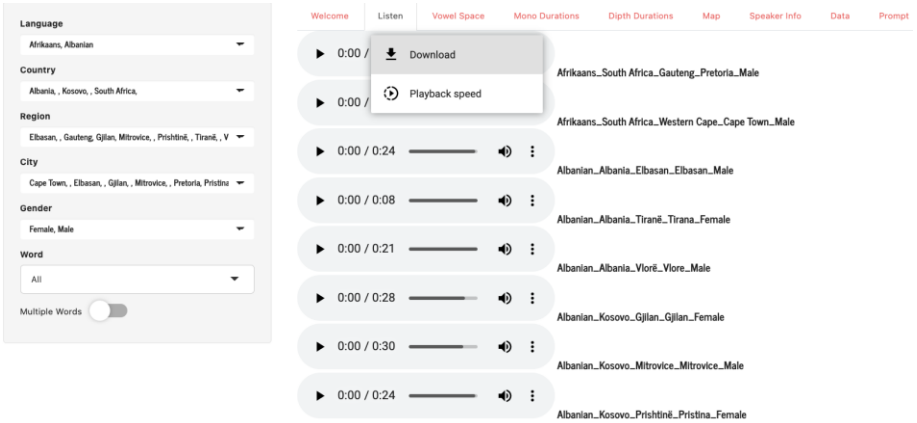


Figure 9. Deployment of app information and content in the Shiny Platform.

5. Conclusions

This paper has demonstrated a methodological approach for building digital materials that can be freely accessed by interested users. We have presented the tools we have used, to show the workflow, from conception to deployment. RStudio and Shiny apps is a combination of platforms that allows us to produce high quality and user-friendly tools that can be maximised by researchers in sociophonetic research. We hope this can be useful for people who may be interested in developing their materials from scratch, using freely available data. Finally, this can be adapted to other studies/material that are not necessarily phonetic in nature, but whose aim is to present digital materials in an online environment.

6. Acknowledgements

I want to thank the anonymous reviewers of this paper for their invaluable comments and insights in the shape and content of the final version. Their generosity and expertise have improved this paper in innumerable ways and saved me from many errors. Those that inevitably remain are entirely my own responsibility.

References

Ahamad, A., Anand, A. & Bhargava, P. (2020). AccentDB: A Database of Non-Native English Accents to Assist Neural Speech Recognition. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, Ch. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, 11–16 May 2020, European Language Resources Association (ELRA), licensed under CC-BY-NC, 5351–5358.

Ahmed, A., Tangri, P., Panda, A., Ramani, D., & Karmakar, S. (2019). VFNet: A Convolutional Architecture for Accent Classification (*2019). IEEE 16th India Council International Conference (INDICON)*, 1-4.

Albin, A. (2014). PraatR: An architecture for controlling the phonetics software "Praat" with the R programming language. *Journal of the Acoustical Society of America* 135(4), 2198.

Amin, T. B., Marziliano, P. & German, J. S. (2014). Glottal and Vocal Tract Characteristics of Voice Impersonators. In *IEEE Transactions on Multimedia*, vol. 16, no. 3, April 2014, 668-678. doi: 10.1109/TMM.2014.2300071

Barreda, Santiago. (2015). *phonTools: Functions for phonetics in R*. R package version 0.2-2.1

Boersma, Paul & Weenink, David (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.16, retrieved 18 August 2022 from http://www.praat.org/

Bořil, T., & Skarnitzl, R. (2016). Tools rPraat and mPraat. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (eds.), *Text, Speech, and Dialogue*. Springer International Publishing, 367–374.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T. & Schnapp, J. (2012). *Digital_Humanities* (PDF). Open

Access eBook: MIT Press. ISBN 9780262312097. Archived from the original (PDF) on 26 October 2016. Retrieved 26 December 2016.
Butryna, A. Chu, S.-C., Demirsahin, I., Gutkin, A., Ha, L., He, F., Jansche, M., Johny, C., Katanova, A., Kjartansson, O., Li, C., Merkulova, T., Oo, Y-M., Pipatsrisawat, K., Rivera, C., Sarin, S., de Silva, P., Sodimana, K., Sproat, R., Wattanavekin, T. & Aris-Eko-Wibawa, J. (2019). Google Crowdsourced Speech Corpora and Related Open-Source Resources for Low-Resource Languages and Dialects: An Overview. In *Proceedings of the Language Technologies for All (LT4All)*, Paris, UNESCO Headquarters, 5-6 December, 2019, 91–94.
Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J. (2019). *shiny: Web Application Framework for R* (Version 1.3.2) [R package]. https://CRAN.R-project.org/package=shiny
Cox, F. (2006). The acoustic characteristics of /hVd/ vowels in the speech of some Australian teenagers. *Australian Journal of Linguistics* 26, 147-179.
Diskin, C., Loakes, D., Billington, R., Stoakes, H., Gonzalez, S. & Kirkham, S. (2019). *The /el/-/æl/ merger in Australian English: Acoustic and articulatory insights.* Melbourne, Australia, 1764-1768.
Docherty, G.-J. & Foulkes, P. (1999). *Derby and Newcastle: Instrumental phonetics and variationist studies*. In Foulkes & Docherty (eds.), Urban Voices. Accent Studies in the British Isles. London: Routledge, 47–71.
Docherty, G., Gonzalez, S. & Mitchell, N. (2015). Static vs Dynamic Perspectives on the Realization of Vowel Nucleii in West Australian English. *Proceedings of the 18th International Congress of Phonetic Sciences International Phonetic Association (IPA).*
Docherty, G.J., Gonzalez, S., Mitchell, N. & Foulkes, P. (2019). An acoustic analysis of short front vowel realizations in the conversational style of young English speakers from Western Australia. *19th International Congress of Phonetic Sciences (ICPhS)*, Melbourne.
Gohel, D., Skintzos, P., Bostock, M. Kokenes, S., Shull, E. & Book, E. (2021). *ggiraph: Create Interactive 'ggplot2'.* Web Application Framework for R (Version 0.7.10) [R package]. https://davidgohel.github.io/ggiraph/
Gordon, E., and Maclagan, M. (2001). Capturing a sound change: a real time study over 15 years of the near/square diphthong merger in New Zealand English. *Australian Journal of Linguistics* 21, 215–238. doi: 10.1080/07268600120080578
Haddican, B., Foulkes, P., Hughes, V., & Richards, H. (2013). Interaction of social and linguistic constraints on two vowel changes in northern England. *Language Variation and Change 25*(3), 371-403. https://doi.org/10.1017/S0954394513000197
Harrington, J. (2010). *Phonetic Analysis of Speech Corpora*. UK: John Wiley & Sons.
Hay, J., Warren, P., and Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *J. Phon*. 34, 458–484. doi: 10.1016/j.wocn. 2005.10.001
Hughes, A., Trudgill, P. & Watt, D. (2005). *English accents and dialects*, 4th edn. London: Hodder Arnold
Jansen, S. & Braber, N. (2020). foot-fronting and foot–strut splitting: vowel variation in the East Midlands. *English Language and Linguistics* 1–31. doi:10.1017/s1360674320000325
Kendall, T. & Thomas, E.-R. (2018). *vowels: Vowel Manipulation, Normalization*. Web Application Framework for R (Version 1.2-2) [R package]. http://blogs.uoregon.edu/vowels/
Labov, W., Karen, M., and Miller, C. (1991). Near-mergers and the suspension of phonemic contrast. *Language Variation and Change* 3, 1, 33–74. doi: 10.1017/S09543945000 00442
Lobanov, B. M. (1971). Classification of Russian vowels spoken by different listeners. *Journal of the Acoustical Society of America* 49, 606–08.
McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Conference of the International Speech Communication Association*, 498–502.
McDougall, K. (2006). Dynamic Features of Speech and the Characterisation of Speakers: Towards a New Approach Using Formant Frequencies. *International Journal of Speech, Language and the Law* 13, 89 126.
Maclagan, M., and Gordon, E. (1996). Out of the air and into the ear: Another view of the New Zealand diphthong merger. *Language Variation and Change* 8, 125–147. doi: 10.1017/S0954394500001095
Mazumderm, M., Ciro, J., Chitlangia, S., Achorn, K., Kanter, D., Diamos, G., Banbury, C., Kang, Y., Galvez, D., Sabini, M., Mattson, P., Warden, P., Meyer, J. & Reddi, V.-J. (2021). Multilingual Spoken Words Corpus. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.*
Nature Portfolio, (2021). *Computational methods*, https://www.nature.com/subjects/computational-methods
Piercy, C. (2011). One /a/ or Two?: Observing a Phonemic Split in Progress in the Southwest of England. University of Pennsylvania Working Papers in Linguistics 17 (2), 18. https://repository.upenn.edu/pwpl/vol17/iss2/18

R Core Team (2021). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/

Rajan, S.-S., Udeshi, S. & Chattopadhyay, S. (2022). AequeVox: Automated Fairness Testing of Speech Recognition Systems. In E. B. Johnsen & M. Wimmer (eds.), *25th International Conference on Fundamental Approaches to Software Engineering (FASE)*, 245–267.

RStudio Team (2021). *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Sievert, C. (2020). Interactive Web-Based Data Visualization with R, plotly, and shiny. *Chapman and Hall/CRC*. ISBN 9781138331457, https://plotly-r.com

Van Heuven, V., J., Edelman, L., van Bezooijen, R. (2002). The pronunciation of /ɜI/ by male and female speakers of avant-garde Dutch. In Broekhuis, H. & Fikkert, P. (eds.), *Linguistics in the Netherlands* 2002, 61–72.

Van Rossum, G., & Drake Jr, F.-L. (1995). Python reference manual. *Centrum voor Wiskunde en Informatica Amsterdam.*

Watson, C., Harrington J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America* 106, 458–468.

Weinberger, S. & Kunath, S. (2015). *Speech Accent Archive*. George Mason University. http://accent.gmu.edu

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.-D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.-L., Miller, E., Bache, S.-M., Müller, K., Ooms, J., Robinson, D., Seidel, D.-P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4 (43), 1686. doi: 10.21105/joss.01686

Wickham, H. (2021). *rvest: Easily Harvest (Scrape) Web Pages*. Web Application Framework for R (Version 1.0.2) [R package]. https://rvest.tidyverse.org/

Wickham, H., François, R., Henry, L. & Müller, K. (2022). *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr

Winkelmann, R., Jaensch, K., Cassidy, S. & Harrington J (2021). *emuR: Main Package of the EMU Speech Database Management System*. R package version 2.3.0.

Zellou, G. & Scarborough, R. (2019). Neighborhood-conditioned phonetic enhancement of an allophonic vowel split. *The Journal of the Acoustical Society of America* 145, 3675–3685.. doi:10.1121/1.5113582

Languages Worldwide and the World Wide Web: Crowdsourcing on the Internet to Explore Linguistic Theories

Mathilde Hutin, Marc Allassonnière-Tang

Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique (UMR 9015)

Muséum National d'Histoire Naturelle, CNRS, Laboratoire d'Eco-Anthropologie (UMR 7206)

E-mail: mathilde.hutin@lisn.upsaclay.fr

Abstract

Vocal languages across the world are estimated to be approximately 6000, yet only a handful of them are well-resourced, thus limiting typological investigations, i.e., language-comparison studies aiming at understanding universal trends in language. Crowd-sourced data could participate in creating homogenous multilingual corpora and therefore provide a revolutionary tool to give researchers access to large amounts of data in rare or remote languages. Yet crowd-sourced data are usually recorded with non-professional tools in non-silent environments, which represents a challenge to anyone wishing to use them for phonetic research. In this paper, we show how crowd-sourced data can participate in academic research by using audio files from Lingua Libre, Wikimedia France's open-access linguistic library, to test the Inventory Size Hypothesis. This hypothesis suggests that the more phonological vowel categories a language has, the less internal phonetic variation vowels will display. The platform allows us to investigate the acoustic measurements of the three cardinal vowels /a/, /i/ and /u/ in 7 less-resourced languages with various numbers of vowel categories. Our results replicate the results of previous literature, which shows that our methodology is promising. Lingua Libre thus successfully allows to investigate a scientific question with theoretical implications for larger models of communication, and to bridge the gap between well and less-resourced languages in an inclusive, homogeneous data set of the world's languages.

Keywords: crowd-sourcing, open-access, linguistic typology, phonetics

1. Linguistics and the Internet: A Virtuous Cycle

Digital Humanities are a new and promising transdiscipline relying on the mutual contribution of digital technologies and human or social sciences. Over the last few decades, one of these human and social sciences – linguistics – has greatly benefitted from both advances in computer sciences and the increasingly generalised access to the Internet, while, conversely, new language technologies were built on the knowledge basis provided by specialists in both theoretical and applied linguistics.

This virtuous cycle however is still frequently restricted to languages for which sufficient data is available (Hovy & Prabhumoye 2021), i.e., to mostly well-documented, well-resourced and unendangered languages. It also causes an imbalance between the various language families of the world. For instance, 78.5% of the Universal Dependencies corpus (Nivre et al. 2019) consists of Indo-European languages. A few tentatives have been made to counter this imbalance, either by creating large multilingual corpora that are more representative of language diversity (cf. *CMU Wilderness Corpus*, Black 2019; *VoxClamantis*, Salesky et al. 2020) or by improving technologies so that they need less data to be trained, thus allowing them to handle less-resourced languages (cf. *Zero Resources Speech Challenge*, Versteegh et al. 2015, Dunbar et al. 2017, 2019, 2020).

Recently, though, this lack of varied data in as many languages as possible has boosted the creation of crowd-sourcing platforms which provide data from virtually all over the world that can be exploited for scientific purposes. Common Voice by Mozilla[47], for instance, gathers crowd-sourced data from volunteers with the intended purpose to provide large corpora to improve speech recognition systems in a number of languages, including less-resourced ones (Ardila et al. 2020). It also allowed the creation of the *VoxCommunis* corpus (Ahn & Chodroff 2022), which provides acoustic models, pronunciation lexicons, and word- and phone-level alignments for 36 languages

[47] https://commonvoice.mozilla.org/fr

from the Common Voice corpus.

In the present paper, we propose to present our exploratory work of how crowd-sourced data can be used for research in linguistics and how, in turn, scientific investigation helps create better tools. We use the data from Lingua Libre[48], an online linguistic library by Wikimedia France which, to the best of our knowledge, was used only once for academic purposes (outside of the present project and related publications), i.e. to estimate the transparency of orthographies in 17 languages using an artificial neural network (Marjou 2021). Contrary to Common Voice, Lingua Libre is not intended to improve speech technologies, or even to provide data for larger scientific research, but to allow patrimonial conservation of languages. It was chosen because it is less normative, and therefore more representative of natural speech, than Common Voice, in which recordings are judged as inadequate in cases of common segmental reduction or hypo-articulation (e.g., French *sûrement*, 'surely', pronounced [syrmɑ̃] instead of [syrəmɑ̃], although the former is more natural than the latter).

However, the creation of these crowd-sourced data sets does not necessarily mean that they can be used for linguistic inquiry. Phonetic studies usually rely on laboratory or field recordings produced by specialists with high-performance recording devices. Yet crowd-sourced data are seldom recorded with professional microphones but rather with computer or even phone-incorporated devices, sometimes in noisy places. Establishing whether such data can indeed be used for phonetic investigation would imply that typological studies can rely on crowd-sourcing to study larger sets of languages, and in particular less-resourced or endangered languages, which is often expensive and difficult to do by traditional means.

In the following, we first describe the theoretical research question we use to test the utility of such crowd-sourced corpora for the scientific community (Section 2). Then we provide information on our corpus and methodology, as well as justify several methodological choices (Section 3). In Section 4, we present our results and, in Section 5, we extensively discuss the outcome of our study and conclude on the advantages of such practices and a few caveats.

2. The Inventory Size Hypothesis

To show how crowd-sourced data can be useful to scientific inquiry, we propose to test it here on a well-known typological question: the Inventory Size Hypothesis.

The Inventory Size Hypothesis (henceforth, ISH) stems from the H&H ("Hypo- and Hyperspeech", Lindblom 1990) communication model, which suggests that, while speaking, speakers constantly oscillate between two functional principles: Reduce the effort for language production on the one hand, which results in less phonetic contrast, and optimise the chances of accurate perception on the other, which results in maximising phonetic contrast. For instance, during speech production, a speaker would be negotiating between (i) hypo-articulating and realising an unspecified, centralised vocoid such as [ə] in lieu of any vowel (e.g., pronouncing something like [bək] for all potential underlying forms *back*, *buck*, etc.), thus needing less effort but risking incomprehension, and (ii) hyper-articulating and realising an overly well-enunciated vowel (e.g., pronouncing [bɑ:k] *vs* [bʌ:k], etc.), thus needing more effort to articulate but diminishing the risk of being misunderstood.

What this negotiation entails, is that the realisation of specific vowels will depend on the number of competitors in the system. For instance, in a language with only three phonological vowels, such as Arabic (see Figure 1a), each vowel has only two competitors, and each can thus expand on a rather large acoustic space, while in a language with 14 vocalic categories, such as German (see

[48] https://lingualibre.org/wiki/LinguaLibre:Main_Page

Figure 1b), each vowel competes with 13 others, and each can only expand to a lesser degree so as to avoid being mixed up with another vowel, which would result in a semantic shift with consequences on the whole message.

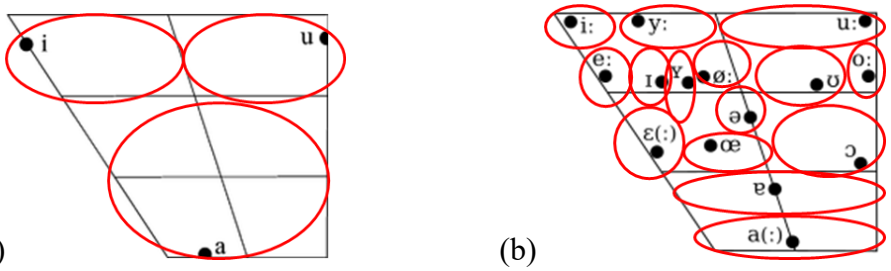


Figure 1: Vowel charts with IPA transcriptions on a two-dimensional vocalic space for the monophthongs of (a) Arabic (source for the background chart: Thelwall 1990:38 cited on Wikipedia) on the left and (b) German (source for the background chart: Dudenredaktion, Kleiner & Knöbl 2015:34 cited on Wikipedia) on the right. The red circles represent the possible variation of each vowel without risking confusion with a competitor.

The consequence, from a typological point of view, would be that the vowels in languages with less vocalic categories will display more internal variation than those from languages with more vocalic categories, and vice-versa, "each vowel act[ing] as a repeller in a dynamical system" (Fletcher & Butcher 2002, 1). Typically, such a hypothesis can be tested only with a reliable set of languages, which is difficult to achieve with professional data, since large homogenous multilingual corpora are rare. The hypothesis has indeed already been tested on professionally gathered data, with contradictory results: A handful of studies tend to validate the hypothesis, but they all rely on a reduced number of languages (cf. Jongman, Fourakis & Sereno 1989 on American English, Greek and German, Al Tamimi & Ferragne 2005 on French and two dialects of Arabic, Peters, Heeringa & Schoormann 2017 on three German languages, and Larouche & Steffann 2018 on Quebec French and Inuktitut). Several other studies tend to invalidate the hypothesis, although also relying on a reduced number of languages (cf. Bradlow 1995 on English and Spanish and Meunier et al. 2003 on English, Spanish and French) that are sometimes even closely related (cf. Recasens & Espinosa 2009 on five dialects of Catalan, Lee 2012 on five dialects of Chinese and Heeringa, Schoormann & Peters 2015 on three German languages). Such studies are, unfortunately, not representative enough to provide either strong support or strong disclaim regarding the ISH.

However, some studies have also taken advantage of the recent technological advances and proposed studies on larger sets on languages (cf. Engstrand and Krull 1991 on 7 languages across 6 language families, Livijn 2000 on 28 languages, Gendrot and Adda-Decker 2007 on 8 languages across 4 families, and Salesky et al. 2020 on 38 languages across 11 families). None of these studies on 7 or more languages from diverse language groups have found evidence for an effect of inventory size on the global acoustic space.

In the present paper, we build on these results and choose to analyse variation in vowel realisation in 7 languages: Afrikaans and German (Germanic), Catalan and Romanian (Romance), Polish and Russian (Slavic) and Basque (isolate). All these languages are considered as less-resourced[49], i.e., the amount of data and technologies to handle them is limited, which means that our study will demonstrate not only that Lingua Libre can be used for phonetic studies, but also that it

[49] See Cieri et al. (2016) for Afrikaans, Polish, Russian and Basque; Burchardt et al. (2012) on German, Moreno et al. (2012) on Catalan, Trandabăt et al. (2012) on Romanian, Miłkowski (2012) and Hutin & Allassonnière-Tang (2022a) on Polish, Hernáez et al. (2012) on Basque.

allows to bridge the gap between well- and less-resourced languages.
The vowel inventories of each language are specified below, ranging from the language with the most vowel categories to the one with the least vowel categories:

- Afrikaans (afr): Afrikaans is a Germanic language mostly spoken in South Africa. It displays 18 vowels: 12 oral /i, y, e, ø, ɛ, œ, a, ɑ, ə, ɔ, o, u/ and 6 nasal /ɪ̃, ʉ̃, ɛ̃, ɑ̃, ɔ̃, ə̃/.
- German (deu): German is a Germanic language mostly spoken in Germany, Austria and Switzerland. It displays 14 vowels: /i, y, ɪ, ʏ, e, ø, ɛ, œ, a, ə, ɔ, o, ʊ, u/.
- Catalan (cat): Catalan is a Romance language mostly spoken in Northern Spain. Central Catalan displays 8 vowels: /i, e, ɛ, a, ə, ɔ, o, u/.
- Romanian (ron): Romanian is a Romance language mostly spoken in Romania. It displays 7 vowels: /i, ɨ, e, ə, a, o, u/.
- Polish (pol): Polish is a Slavic language mostly spoken in Poland. It displays 6 oral vowels: /i, ɪ, ɛ, a, ɔ, u/ and 2 nasals /ẽ, õ/.
- Russian (rus): Russian is a Slavic language mostly spoken in Russia. It displays 6 vowels: /i, ɪ, e, a, o, u/.
- Basque (eus): Basque is an isolate mostly spoken in Northern Spain and South-Western France. It displays 5 vowels: /i, e, a, o, u/[50].

The choice of languages for this study is explained below (see Section 3).

3. Material and Method

To test our methodology on so many languages, we use the data from Lingua Libre. Lingua Libre is Wikimedia France's linguistic library, aiming to counter the lack of oral data as well as of diversity in the languages represented on the Internet. Since its launch in 2015, it has, to this date (August 2022) gathered ~710k short recordings (either isolated words or short expressions) in 159 languages across 857 speakers. As a crowd-sourcing tool relying on Wikimedia's philosophy, any speaker can log into Lingua Libre, fill in a profile with basic metadata (pseudonym, gender, languages and proficiency, geographical localization and choice of licence), and record themselves (or guests) reading lists of words or of chunks of words, either in their native language or in a second language. The device detects pauses, which allows for the recording to end when the word or expression has been read and the next recording to start automatically after, therefore effortlessly generating relatively short audio files for each utterance. Each audio file is supposed to be titled on the same template of 'Language - Speaker - Item'. For example, the recording titled "pol.-KaMan-dokumentalny.wav" comprises an audio file in Polish ("pol."), spoken out by "KaMan", and the recorded item is "dokumentalny", which means "documentary". All audio files are open-source, i.e., under a Creative Commons licence.

Regarding the methodology, the recordings are first scraped from the Lingua Libre database and then segmented and aligned using WebMAUS (Kisler, Reichel & Schiel 2017). WebMAUS is the online open-access version of the Munich AUtomatic Segmentation (MAUS) software (Schiel 1999; 2004), which is used to automatically align a recording based on its orthographic transcription. To do so, it creates a pronunciation hypothesis graph (several potential transcriptions and alignments) in the Speech Assessment Methods Phonetic Alphabet (SAMPA, Wells 1997) based on the orthographic transcript of the recording using a grapheme-to-phoneme converter. The signal is then aligned with the hypothesis graph and the alignment with the highest probability is chosen. This tool has been shown to display a 95% accuracy compared to manual alignments (Kipp, Wesenick & Schiel 1997).

[50] These inventories are based on the SAMPA inventories used by the MAUS software. Some inaccuracies can emerge from this choice, since classical grammars sometimes provide different inventories (see Phoible 2.0).

Once the audio has been aligned with a SAMPA transcription, the selected vowels are extracted from the recordings. In the present study, we investigate only the three cardinal vowels /a, i, u/, which are supposedly the three universal vowels shared by all vocal languages (Crothers 1978, Schwartz et al. 1997a; b).

Finally, the extracted recordings of the selected vowels are analysed in terms of formants. Formants are peaks of energy, measured in Hertz, that allow to distinguish sounds, as can be seen in red on the spectrograms in Figure 2a-c for Romanian [a] vs [i] vs [u].

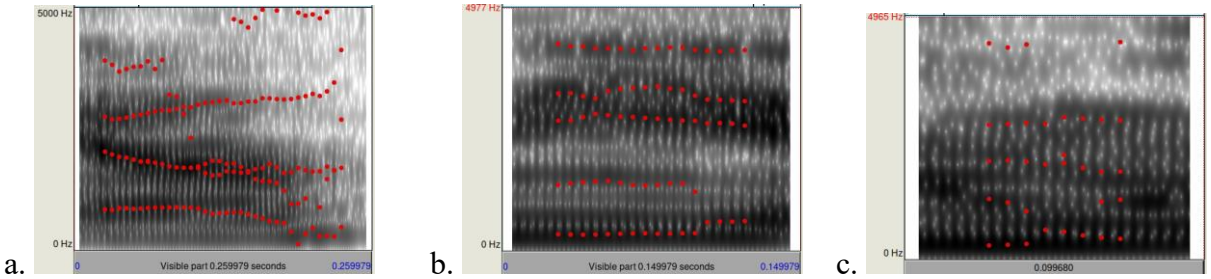


Figure 2: Spectrograms for (a) the vowel [a], (b) the vowel [i] and (c) the vowel [u] pronounced by a Romanian speaker. The red dotted lines allow to make the five formants visible. The lowest two, aka F1 and F2, are different for each vowel-type, but usually similar across vowel-tokens.

The first two formants (out of five) are traditionally used to discriminate vowels, as can be seen in Figure 3. The three cardinal vowels /a/, /i/ and /u/ are the most peripheral ones: /a/ is usually realised with an F1 value around 900 Hz and an F2 value of approximately 1600 Hz, while /i/ and /u/ have very low F1 values (usually around 200 or 300 Hz), with a high F2 value of 2400 Hz for /i/ and a low one of 600 Hz for /u/.

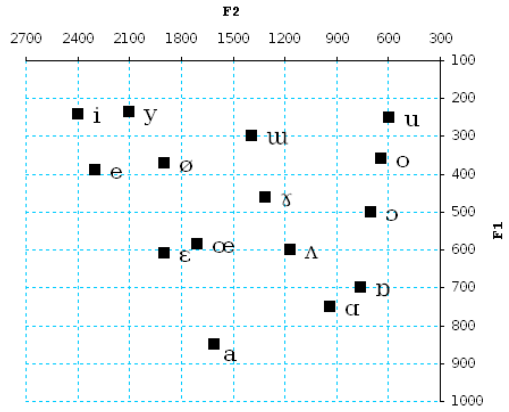


Figure 3: Vocalic chart with corresponding average values for F1 and F2 (Source: Любослов Езыкин, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons).

For each recording of each vowel, the F1 and F2 at the middle of the entire sound are extracted. This step is expected to diminish the influence of context-induced noise in the recordings. During this process of data extraction and analysis, the following R packages are used: emuR (Winkelmann, Harrington & Jänsch 2017), PraatR (Albin 2014), and tidyverse (Wickham 2017). The code and the data used for the analysis are available in the supplementary materials.

In the present study, we extract a subsample of 50 items for each vowel in each of the 7 languages, in order to counter the fact that some languages have much more data points than others.

The threshold 50 is decided based on the minimum count of the targeted vowels in a sample of 1000 recordings for each of the selected languages. We also aim at having a similar number of speakers (10) for each language. In total, we thus investigated 50 items * 3 vowels * 7 languages, i.e., 1050 vowel utterances.

The 7 languages were chosen on a number of technical factors:

1. They have at least 1000 recordings in Lingua Libre, so that we have access to a substantial sample,
2. They do not use tones as a distinctive feature, since tones can add complexity to the alignment process and to formant analysis,
3. They do not use vowel harmony (i.e., levelling in the quality of all the vowels in a word), which could hide the underlying nature of the vowel,
4. They are manageable in MAUS and they have the three vowels /a, i, u/ in their MAUS inventory (for instance, English had enough data in Lingua Libre and a segmentation tool in MAUS, but the vowel /a/ was not taken into account by the MAUS inventory for English, which can only align /æ/ and /ɑ/).

As a result, we obtain the data displayed in Table 1.

Language	ISO	(Oral) vowel inventory[51]	[a]	[i]	[u]	Number of speakers
German	deu	14	551	225	70	13
Afrikaans	afr	12	228	249	87	8
Catalan	cat	8	398	478	325	9
Romanian	ron	7	794	524	377	13
Polish	pol	6	936	347	162	14
Russian	rus	6	951	560	199	14
Basque	eus	5	1376	542	253	14

Table 1: The distribution of the vowels [a], [i], and [u] in a sample of 1000 recordings for each language from Lingua Libre. The vowel inventory is the number of timbres in each language.

Finally, the data set used for the statistical analyses of Section 4 looks like the sample displayed in Table 2. The data set displays, first, the vowel category, then an ID number specific to each vowel-token, third the ISO-code of the language, then the values for F1 and F2 as measured in the middle of the vowel-token, the ID of the speaker and finally the word in which the vowel was pronounced.

Vowel	ID	ISO	F1	F2	Speaker	Item
a	1871	cat	673	2453	Unjoanqualsevol	preguntar
i	39	afr	377	2268	Anon1314	fermium
u	3548	eus	369	2113	Xabier_Cañas	adur

Table 2: A sample of the data extracted from Lingua Libre. The rows are occurrences of the vowels [a], [i], and [u] in languages of the data set.

This data allows us to map vocalic values on a two-dimensional space, as exemplified in Figure 4.

[51] Since nasality has particular acoustic correlates such as anti-formants, we consider nasal vowels as not confusable with oral vowels and focus solely on the latter.

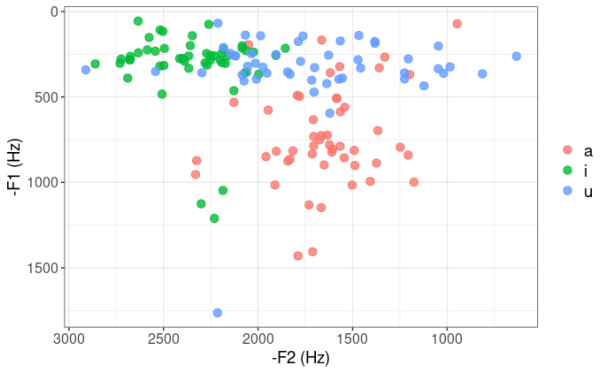


Figure 4: Formants for occurrences of the vowels [a], [i], and [u] in the Romanian recordings data set extracted from Lingua Libre. The formants are extracted from the central point for each occurrence of each vowel.

Based on the visualisation in Figure 4, we acknowledge that there are obviously a few outliers in the extracted occurrences. For instance, some occurrences of [a] have very different formants from the other occurrences and appear to resemble /ə/ or even /ɨ/, and some occurrences of [u] seem to be closer to /ɨ/ or /y/. This is likely due to the segmentation accuracy, as some parts of the surrounding context, i.e., the left and right segments, might have been included by the automatic segmentation.

4. Results: Lingua Libre Data Invalidates the Inventory Size Hypothesis

In this section, we provide the results of several plotting techniques and statistical analyses to test our hypothesis, i.e., whether the size of the vocalic inventory influences the internal variation in the realisation of each vowel.

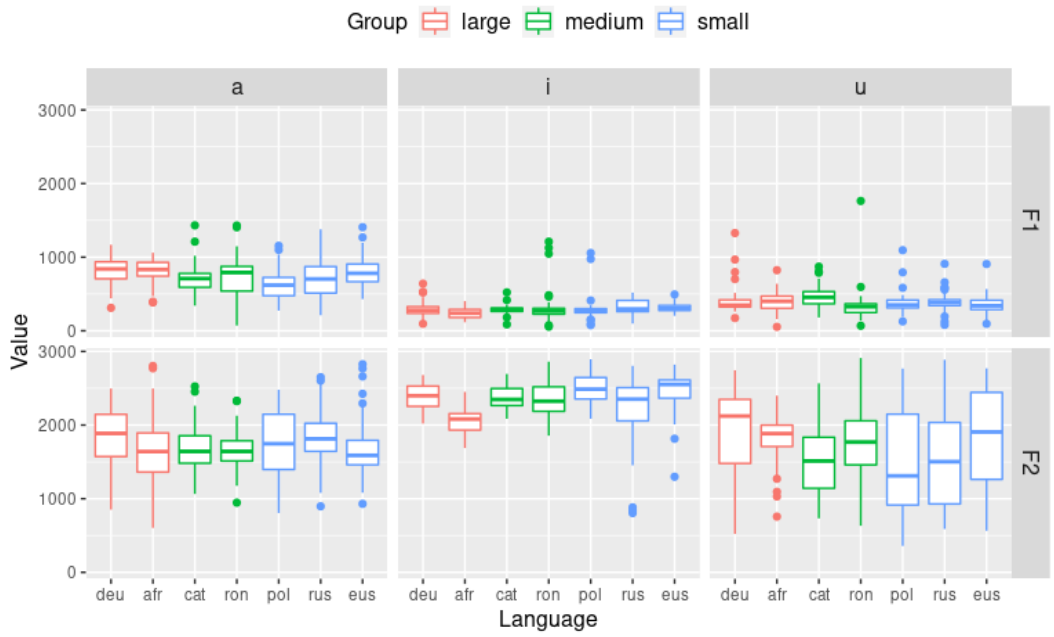


Figure 5: The distribution of formants for [a], [i] and [u] across the 7 languages of the data set. The languages on the x-axis are ranked according to their number of timbres from high to low. The colours indicate the group to which the language is affiliated with in terms of number of timbres: red = large, green = medium, blue = small.

Figure 5 plots the distribution of the first two formants (F1 on the top-tier and F2 on the bottom-tier) for each vowel (/a/, /i/ and /u/ from left to right): The bigger the box, the more variation has been observed. Visually, the boxes are not generally bigger for languages with small vowel inventories (5

or 6 vowels), except maybe for the F2 of /u/.

Figure 6 shows the variation of F1 (top-tier) and F2 (bottom-tier) of each vowel, when grouping the languages with regards to the size of their inventory, i.e., when assigning languages to the group of "large" inventories (arbitrarily defined as having more than 10 vocalic categories, i.e., Afrikaans and German), "medium" inventories (arbitrarily defined as having less than 10 but more than 6 vowel categories, i.e., Catalan and Romanian) or "small" inventories (arbitrarily defined as having 6 or less vowel categories, i.e., Polish, Russian and Basque). Here the y-axis does not indicate raw formant values in Hertz anymore, but the standard deviation, i.e., the higher the position of the box, the higher the variation.

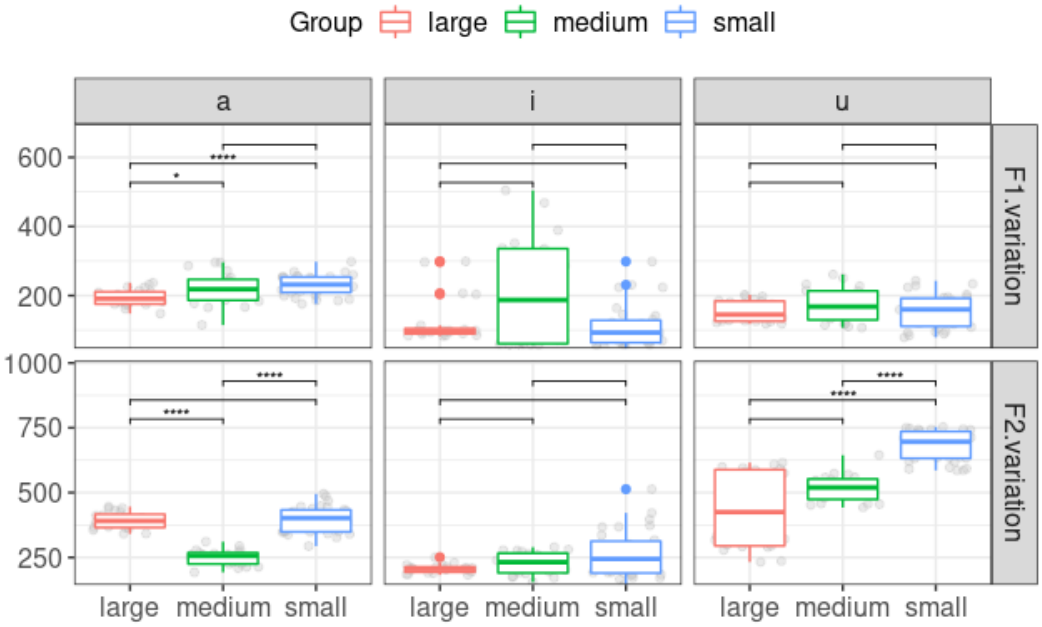


Figure 6: Distribution of the standard deviation of formants based on inventory groups for [a], [i], and [u] extracted from the Lingua Libre data across ten replications. The significance labels indicate the output of a Wilcoxon test with Bonferroni correction (with adjusted significance stars, where * means p < = 0.05, ** means p < = 0.01, and *** indicates p < = 0.001. Unmarked brackets show that the test is not significant).

Results show that there is a statistically significant difference mainly for [a] in both F1 and F2 values, especially between the F1 standard deviation of languages with a large *vs* a small vowel inventory, but the difference in F2 standard deviation is not statistically significant between these two groups, and the pattern is not regularly increasing, as would have been expected, with languages with medium inventories generally displaying much less variation in F2 than the other two groups. The three language groups however do not vary in a statistically significant manner for the realisation of [i], be it in F1 or in F2. For [u], only the realisation of F2 differs significantly between languages with a large or a medium *vs* a small inventory, but not between a large and a medium inventory.

As an additional exploration, we also consider the variation of the acoustic space covered by each vowel category in each language. Taking the content of Figure 4 as an example, we use a 2D kernel density estimation (Venables & Ripley 2002) to extract the contour of the area covered by each vowel in the formant space. We set the number of contours to 4 to remove outliers in the data that may overextend the coverage of a vowel in the acoustic space. Taking the vowel [a] in Romanian as an example (Figure 7), we can extract the area of the contour covered by the occurrences of [a] in the data.

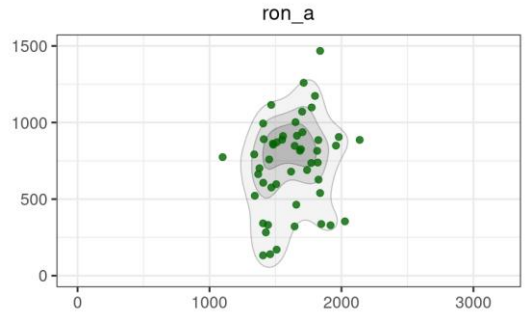


Figure 7: Distribution of formants for the vowel [a] in Romanian based on an extraction of 50 occurrences. The x-axis represents F2 and the y-axis indicates F1. The contours are generated with a 2D kernel density estimation.

To estimate whether the number of oral vowels in the phonemic inventory correlates with the variation in the realisation, we conducted a linear mixed model with the standard deviations of F1 and F2 along with areas in the acoustic space as the dependent variables. The results of this statistical analysis can be seen in Table 3.

Dependent variable	Predictor	Estimate	df	t value	p value
F1_sd	nb_of_timbers	-1.341	5	-0.267	0.800
	vowel_i	-75.704	201	-7.390	p < 0.001 ***
	vowel_u	-58.251	201	-5.686	p < 0.001 ***
F2_sd	nb_of_timbers	-8.915	5	-1.141	0.305
	vowel_i	-118.803	201	-8.788	p < 0.001 ***
	vowel_u	209.522	201	15.499	p < 0.001 ***
Area	nb_of_timbers	-18952.723	5	-2.621	0.047 *
	vowel_i	-394406.179	201	-11.689	p < 0.001 ***
	vowel_u	-65096.748	201	-1.929	0.055

Table 3: Output of linear mixed models based on the output of 10 vowel samplings with 50 tokens for each vowel in each language. The abbreviations are read as follows: nb_of_timbers = number of timbers, sd = standard deviation, df = degrees of freedom

As can be seen in Table 3, the results for F1 are not statistically significant ($p = 0.8$). However, those for F2 are significant ($p = 0.3$), suggesting that at least one dimension in the acoustical values of vowels may be impacted by the phonemic inventory. However, when looking at the general area covered by the vowels on the acoustic space, the difference is much smaller ($t = -2.621$) and the result barely significant ($p = 0.047$).

These results generally indicate that the ISH is not completely ruled out, but that it is unlikely the most accurate hypothesis to account for vocalic variation. Further studies should also control for roundness, which typically lowers formants and implies an important role of F3 (Vaissière 2011). A preliminary study (Hutin & Allassonnière-Tang 2022b) also suggests that it is not the size, but the shape of the inventory, that influences the scope of variation.

5. Conclusion and discussion

In this paper, we proposed to show how crowd-sourced data can participate in academic research. To that extent, we used this kind of data, i.e., audio files from Lingua Libre, to test a scientific hypothesis suggesting that the more phonological vowel categories a language has, the less phonetic internal variation vowels will display. To show that our methodology is promising, we hoped to replicate

results from past research, almost all invalidating the hypothesis. Our expectations were met, as none of our plotting methodologies nor statistical analyses provided strong support for the hypothesis. Lingua Libre thus successfully allowed us to investigate a scientific question with theoretical implications for our models of communication, and to bridge the gap between well and less-resourced languages.

Our study shows that Lingua Libre has an impressive potential. It provides great amounts of data in many languages, the audio recordings show reliable quality and the documentation to access and exploit it is easily accessible and usable. Moreover, as an open-access, crowd-sourced tool, it has the advantage of offering an ever-growing data set.

However, we also came across several difficulties, in particular regarding the metadata. First, linking the data to the speakers' metadata proved difficult, since the process has not yet been automatised in Lingua Libre, and we had to process it manually. Moreover, the metadata was not always accurately filled by participants. Our suggestions for improvement would thus be to improve the processing of the metadata. This implies not only facilitating the linking between files and metadata, but also raising the contributors' awareness on the need for accurate, clean information. A verification tool *à la* Common Voice could also improve the quality of the recordings, provided it stays true to Wikimedia's philosophy of diversity by allowing, contrary to Common Voice, more varied pronunciations as adequate variants of a word.

Since Lingua Libre proved useful here to investigate a specific research topic, we hope to use it again for future research. On this same research topic, we would like to investigate more linguistic variables (such as the duration and F3, F4, and F5 of the vowels), as well as sociolinguistic variables (such as gender, geographical location, etc.). But most of all, we hope that the data set will continue to grow so that we can include more and more languages to our study.

6. Acknowledgements

This research was partially supported by the Excellency Award of Institut DATAIA and the MSH Paris-Saclay (grant OTELO: OnTologies pour l'Enrichissement de l'analyse Linguistique de l'Oral, PI Ioana Vasilescu and Fabian Suchanek) and by the French National Research Agency (grant EVOGRAM: The role of linguistic and non-linguistic factors in the evolution of nominal classification systems, ANR-20-CE27-0021, PI Marc Allassonnière-Tang).

References

Ahn, E.P. & Chodroff, E. (2022). VoxCommunis: A Corpus for Cross-linguistic Phonetic Analysis. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*. Language Resources and Evaluation Conference (LREC 2022).

Albin, A. (2014). PraatR: An architecture for controlling the phonetics software "Praat" with the R programming language. *Journal of the Acoustical Society of America* 135 (4), 2198–2199.

Al-Tamimi, J.E. & Ferragne, E. (2005). Does vowel space size depend on language vowel inventories? Evidence from two Arabic dialects and French. In *Proceedings of Interspeech Eurospeech 2005*, 2465–2468.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M. & Weber, G. (2020) Common Voice: A Massively-Multilingual Speech Corpus. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. & Piperidis, S. (eds.), *Proceedings of LREC*, 4218–4222.

Black, A. W. (2019). CMU Wilderness Multilingual Speech Dataset. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 5971–5975. http://doi.org/10.1109/ICASSP.2019.8683536.

Bradlow, A.R. (1995). A comparative acoustic study of English and Spanish vowels. *Journal of the Acoustical*

Society of America 97, 1916–1924. https://doi.org/10.1121/1.412064
Burchardt, A., Egg, M., Eichler, K., Krenn, B., Kreutel, J., Leßmöllmann, A., Rehm, G., Stede, M., Uszkoreit, H. & Volk, M. (2012). *Die Deutsche Sprache im Digitalen Zeitalter – The German Language in the Digital Age*. META-NET White Paper Series. Berlin: Springer. http://www.meta-net.eu/whitepapers/volumes/german
Cieri, C., Maxwell, M., Strassel, S. & Tracey, J. (2016). Selection Criteria for Low Resource Language Programs. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis., S., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia, 23–28 May 2016. European Language Resources Association (ELRA), 4543–4549. https://aclanthology.org/L16-1720
Crothers, J. (1978). Typology and universals of vowel systems. In J. H. Greenberg, C. A. Ferguson, & E. A. Moravcsik (eds.), *Universals of Human Language*. Vol. 2: Phonology, 93–152. Stanford: Stanford University Press.
Dudenredaktion; Kleiner, S. & Knöbl, R. (2015) [First published 1962], *Das Aussprachewörterbuch* (7th ed.), Berlin: Dudenverlag. ISBN 978-3-411-04067-4
Dunbar, E., Cao, X.N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X. & Dupoux, E. (2017). *The Zero Resource Speech Challenge 2017*. http://doi.org/10.1109/ASRU.2017.8268953
Dunbar, E., Algayres, R., Karadayi, J., Bernard, M., Benjumea, J., Cao, X.N., Miskic, L., Dugrain, C., Ondel, L., Black, A.W. Besacier, L., Sakti, S. & Dupoux, E. (2019). The Zero Resource Speech Challenge 2019: TTS Without T. In *Proceedings of Interspeech 2019*, Graz Austria, 15–19 September 2019, 1088–1092. http://doi.org/10.21437/Interspeech.2019-2904
Dunbar, E., Karadayi, J., Bernard, M., Cao, X.N., Algayres, R., Ondel, L., Besacier, L., Sakti, S. & Dupoux, E. (2020). The Zero Resource Speech Challenge 2020: Discovering Discrete Subword and Word Units. In *Proceedings of Interspeech 2020*, 4831–4835. http://doi.org/10.21437/Interspeech.2020-2743
Engstrand, O. & Krull, D. (1991). Effects of inventory size on the distribution of vowels in the formant space: preliminary data from seven languages. *PERILUS*, 15–18.
Fletcher, J. & Butcher, A. (2002). Vowel dispersion in two northern Australian Languages: Dalabon and Bininj Gun-wok. In C. Bow (ed.), *Proceedings of the 9th Australian International Conference on Speech Science and Technology*, 343–348. Melbourne: Australian Speech Science and Technology Association.
Gendrot, C. & Adda-Decker, M. (2007). Impact of duration and vowel inventory on formant values of oral vowels: An automated formant analysis from eight languages. *International Conference on Phonetics Sciences*, 1417–1420.
Heeringa, W., Schoormann, H. & Peters, J. (2015). Cross-linguistic vowel variation in Saterland: Saterland Frisian, low German, and high German. *Journal of the Acoustical Society of America*, 25–29.
Hernáez, I., Navas, E., Odriozola, I., Sarasola, K., Diaz de Ilarraza, A., Leturia, I., Diaz de Lezana, A., Oihartzabal, B. & Salaberria, J. (2012). *Euskara Aro Digitalean – The Basque Language in the Digital Age.* META-NET White Paper Series. Berlin: Springer. http://www.meta-net.eu/whitepapers/volumes/basque
Hovy, D. & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass* 15 (8). https://doi.org/10.1111/lnc3.12432
Hutin, M. & Allassonnière-Tang, M. (2022a). Crowd-sourcing for Less-resourced Languages: Lingua Libre for Polish. In *Proceedings of SIGUL 2022*.
Hutin, M. & Allassonnière-Tang, M. (2022b). Investigating phonological theories with crowd-sourced data: The Inventory Size Hypothesis in the light of Lingua Libre. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology.* Association for Computational Linguistics (ACL).
Jongman, A., Fourakis, M. & Sereno, J.A. (1989). The Acoustic Vowel Space of Modern Greek and German. Language and Speech 1989, 32, 221–248. http://doi.org/10.1177/002383098903200303
Kipp, A., Wesenick, M.-B. & Schiel, F. (1997) Pronuncation modeling applied to automatic segmentation of spontaneous speech. In *Proceedings of the 5th European Conference on Speech Communication and Technology* (Eurospeech 1997), 1023-1026
Kisler, T., Reichel, U. & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language* 45, 326–347. https://doi.org/10.1016/j.csl.2017.01.005
Larouche, C. & Steffann, F. (2018). Vowel space of French and Inuktitut: An exploratory study of the effect of vowel density on vowel dispersion. In *Proceedings of the Workshop on the Structure and Constituency of Languages of the Americas*; of British Columbia Working Papers in Linguistics 46., U., Ed., Vol. 21.
Lee, W.S. (2012). A cross-dialect comparison of vowel dispersion and vowel variability. *8th International Symposium on Chinese Spoken Language Processing*, 25–29.
Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In Hardcastle, W.J.; Marchal, A. (eds.), *Speech Production and Speech Modelling*. Springer Netherlands: Dordrecht, 403–439.

http://doi.org/10.1007/978-94-009-2037-8_16
Livijn, P. (2000). Acoustic distribution of vowels in differently sized inventories - hot spots or adaptive dispersion? *PERILUS*, 93–96.
Marjou, X. (2021). OTEANN: Estimating the Transparency of Orthographies with an Artificial Neural Network. *Association for Computational Linguistics*. http://doi.org/10.18653/v1/2021.sigtyp-1.1
Meunier, C., Frenck-Mestre, C., Lelekov-Boissard, T. & Le Besnerais, M. (2003). *Production and perception of vowels: does the density of the system play a role?* Université Autonome de Barcelone, 723–726.
Miłkowski, M. (2012) *Język polski werze cyfrowej – The Polish Language in the Digital Age.* META-NET White Paper Series. Berlin: Springer. http://www.meta-net.eu/whitepapers/volumes/polish
Moreno, A., Bel, N., Revilla, E., Garcia, E. & Vallverdú, S. (2012). *La llengua catalana a l'era digital. – The Catalan Language in the Digital Age.* META-NET White Paper Series. Berlin: Springer. http://www.meta-net.eu/whitepapers/volumes/catalan
Nivre, J., Abrams, M., Agic, Z.; Ahrenberg, L., Aleksandravičiūtė, G., Antonsen, L., Aplonova, K., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu M., … Zhu, H. (2019). *Universal Dependencies 2.4*; LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL): Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2988.
Peters, J., Heeringa, W. J. & Schoormann, H. E. (2017). Cross-linguistic vowel variation in trilingual speakers of Saterland Frisian, Low German, and High German, *Journal of the Acoustical Society of America 142*, 991–1005 https://doi.org/10.1121/1.4998723
Recasens, D. & Espinosa, A. (2009). Dispersion and variability in Catalan five and six peripheral vowel systems. *Speech Communication* 51, 240–258. http://doi.org/10.1016/j.specom.2008.09.002
Salesky, E., Chodroff, E., Pimentel, T., Wiesner, M., Cotterell, R., Black, A.W. & Eisner, J. (2020). A Corpus for Large-Scale Phonetic Typology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics*. Online, 4526–4546. http://doi.org/10.18653/v1/2020.acl-main.415
Schiel, F. (1999). Automatic Phonetic Transcription of Non-Prompted Speech. In Ohala, J. J. (ed.), *Proceedings of ICPhS*, 607–610.
Schiel, F. (2004): MAUS Goes Iterative. In Lino, M. T., Xavier, M. F., Ferreira, F., Costa, R. & Silva, R. (eds.). *Proceedings of the LREC 2004*, 1015–1018.
Schwartz, J.-L., Boe, L.-J., Vallee, N. & Abry, C. (1997a). Major trends in vowel system inventories. *Journal of Phonetics* 25 (3), 233–253
Schwartz, J.-L., Boë, L.J., Vallée, N. & Abry, C. (1997b). The Dispersion-Focalization Theory of vowel systems. *Journal of Phonetics* 25, 255–286. http://doi.org/10.1006/jpho.1997.0043
Trandabăt, D., Irimia, E., Barbu Mititelu, V., Cristea, D. & Tufis, D. (2012). *Limba română în era digitală – The Romanian Language in the Digital Age*. META-NET White Paper Series. Berlin: Springer. http://www.meta-net.eu/whitepapers/volumes/romanian
Thelwall, R. (1990). Illustrations of the IPA: Arabic, *Journal of the International Phonetic Association*, 20 (2), 37–41. http://doi.org/10.1017/S0025100300004266
Vaissière, J. (2011). On the acoustic and perceptual characterization of reference vowels in a cross-language perspective. In Zee, E. (ed). *Proceedings of the 17th International Congress of Phonetic Sciences* (ICPhS XVII). China, Aug 2011, 52–59. <halshs-00676266>
Venables, W.N. & Ripley, B.D. (2002). *Modern applied statistics with S*. New York: Springer.
Versteegh, M., Thiollière, R., Schatz, T., Cao, X.N., Anguera, X., Jansen, A. & Dupoux, E. (2015). The zero resource speech challenge 2015. In *Proceedings of Interspeech 2015*. Dresden Germany, 6–10 September 2015, 3169–3173. http://doi.org/10.21437/Interspeech.2015-638
Wells, J. (1997). SAMPA computer readable phonetic alphabet. In Gibbon, D., Moore, R., Winski, R. (eds.). *Handbook of Standards and Resources for Spoken Language Systems.* Part IV. Mouton de Gruyter.
Wickham, H. (2017). tidyverse: Easily install and load the Tidyverse. *R package version 2017*, 1.2.1.
Winkelmann, R., Harrington, J. & Jänsch, K. (2017). EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language*, Volume 45, 392–410.

Use of Sign Language Videos in EEG and MEG Studies. Experiences from a Multidisciplinary Project Combining Linguistics and Cognitive Neuroscience

Doris Hernández,[1] Anna Puupponen,[1] Jarkko Keränen,[1] Tuija Wainio,[1] Outi Pippuri,[1] Gerardo Ortega,[2] Tommi Jantunen[1]

[1]Sign Language Center, Department of Language and Communication, University of Jyväskylä

[2]University of Birmingham

E-mail: doris.m.hernandez-barros@jyu.fi

Abstract

In this paper, we describe our experiences of bringing together methodologies of two disciplines – sign language (SL) linguistics and cognitive neuroscience – in the multidisciplinary ShowTell research project (Academy of Finland 2021–2025). More specifically, we discuss the challenges we encountered when creating and using video materials for the study of SL processing in the brain. Rather than using still images, the study of SL comprehension is better performed by using videos, thus providing more naturalistic stimuli as observed in face-to-face interaction. On the other hand, in neuroimaging (electroencephalography [EEG]/magnetoencephalography [MEG]), it is vital to track the timing of the stimulation exactly and to minimize the noise that could arise from inside and outside the brain. Any brain activity not related to the specific aspect being studied could create artifacts that diminish the signal-to-noise ratio of the measurements, thus compromising the quality of the data. This creates significant challenges when integrating both disciplines into the same study. In the paper, we (i) describe the process of, and requirements for, creating signed video materials that try to mirror naturalistic signing; (ii) discuss the problems in the synchronization of the video stimuli with the brain imaging data; and (iii) introduce the steps we have taken to minimize these challenges in different phases of the process, such as the design, recording, and processing of the video stimuli. Finally, we discuss how, with the use of these steps, we have been able to deal successfully with the resulting data and creating materials that integrate the naturalistic nature of human communication.

Keywords: sign language, multidisciplinary, neurolinguistics, video stimuli

1. Background

In sign languages (SL), linguistic messages are visually captured based on movements of the hands and other parts of the body. This makes studying SLs different from studying spoken (oral-auditory) languages: the difference in the modality requires different methodologies in parts, although ultimately, both languages may be connected to the same biological and cognitive underpinnings. Historically, SLs are not directly related to spoken languages, but in the modern and global world, most SLs are in contact with the surrounding spoken languages of ambient society. Today, SLs are not only used by deaf people, but also by hearing native signers and hearing second language learners who are in contact with deaf societies. Furthermore, many deaf people sometimes use many languages, both other sign languages and/or written languages.

The study of languages such as SLs is a complex process that involves diverse perspectives such as linguistic, cognitive, and social viewpoints. Thus, it might be better approached in a multidisciplinary manner by bringing together the diverse approaches of different disciplines. But combining two distinct disciplines is not always easy due to differences in research histories, paradigms, methodologies, concepts, theoretical starting points, etc. This has been experienced in the research project "ShowTell – Showing and telling in Finnish Sign Language" (jyu.fi/showtell), funded by the Academy of Finland 2021–2025. The ShowTell project investigates how showing meaning is connected to telling meaning in Finnish Sign Language (FinSL) by analyzing the relationship between bodily enactment and traditional language use with lexical items. The project addresses this issue by approaching it from the perspectives of language use (FinSL corpus data, Study 1), kinematic movement production (motion capture data, Study 2), and brain-based meaning processing (functional neuroimaging data, Study 3). During the first year of the project, the research

material for the Study 3 (signed sentences and individual signs) was built and tested. The content of the current paper is based on these processes.

During the creation of the research material, the combination of two related but distinct disciplines – SL linguistics and cognitive neuroscience – resulted in significant challenges for the completion of this stage (study planning and the creation of stimuli). Although linguistics and cognitive neuroscience might study the same phenomenon, they might also use different traditions and paradigms that can impose the use of conflicting requirements on the stimulus materials to be used. How, then, can the multidisciplinary integration of both methods be approached for a more integrated study of SLs? The aim of this article is to document the experiences of combining SL video materials and cognitive neuroscience methodologies. We would like to highlight that this is not a report of an original research, but rather an article where we document our experiences. The specific challenges we have encountered will be described in the next sections as well as the solutions we have implemented for each of them. Finally, we provide an example of testing the implementations in the pilot results.

2. Doing SL Linguistics from a Cognitive Neuroscience Perspective

At the beginning of the modern study of SLs in the 1960s and 70s, analysis was restricted to observing unrecorded SL use, still images, or textual descriptions of signs. As video technologies have advanced, studies on different aspects of SLs have expanded. The analysis of SL structure and use, for example, has unprecedented possibilities due to the accessibility of high-quality recording equipment and the new corpus infrastructures that have been built during the last decades (Orfanidou, Woll & Morgan 2015; Salonen, Kronqvist & Jantunen 2020). This development is essential for the study of visual - gestural languages. To understand how SLs are produced, what they are, and how they work, analysis needs to be done based on formats, such as video, that capture the characteristics of the three-dimensional and multimodal articulation of signers as effectively as possible. In data-driven work on different fields of SL studies, this is a rather self-evident fact – a starting point, so to speak.

The same also applies to the study of SL processing. To understand the perception and cognitive processing of an SL, the stimuli should be as close as possible to the ways in which SLs are perceived in real life. SL do not have a writing system so psycholinguistic and neurocognitive studies have to resort to videos. While using still images or videos of single lexical signs as stimuli is a good starting point and ensures that the stimulation is as controlled as possible, our understanding from the development of the field (Hernández, Puupponen & Jantunen 2022) is that we should go more towards naturalistic signing in creating the stimuli. This includes, for example, the use of signed sentences recorded on the video.

In cognitive neuroscience, brain signals are usually recorded (with electroencephalography [EEG] or magnetoencephalography [MEG]) simultaneously, while the participants solve a cognitive task or process stimuli that are specifically designed. Cognitive neuroscience paradigms/tasks tend to compare different types of stimuli under different conditions. Later, differences (such as amplitude or latency) in each component or time window might reflect processing differences. This allows researchers to identify functional brain-related markers for processing a task or stimulation (see Figure 1). EEG and MEG rely on synchronized electrical activity (and the magnetic field arising from it) coming from large groups of neurons with a specific orientation with respect to the scalp (Luck 2014). Electricity is conducted swiftly; thus, the signal is particularly sensitive to changes that occur over time.

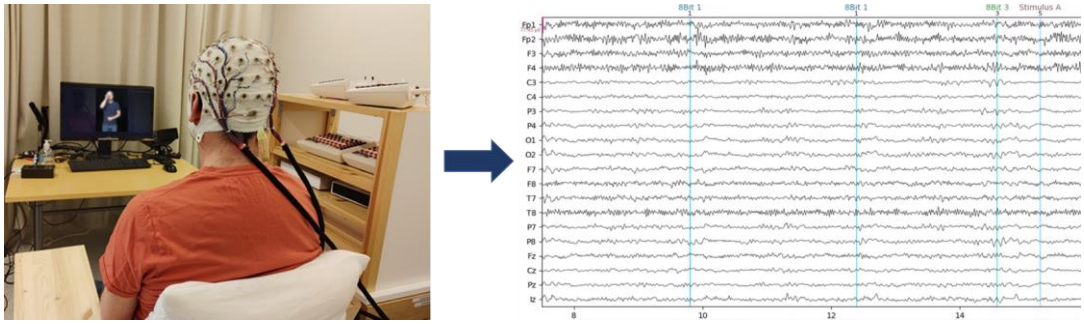


Figure 1. Schematic representation of the EEG recording process.

In EEG and MEG recordings, the temporal characteristics of the stimulus require particular attention. Every change in the stimulus that can influence neurocognitive processing should be properly and accurately marked (in the order of milliseconds [ms]) to allow its link with brain data. Different stimuli elicit different responses (components), whose amplitudes might reflect the degree of neural engagement in different cognitive processes. Furthermore, the stimuli to be used need to be carefully created. Frequently, simplified stimulation (such as geometric figures for visual or simple tones for auditory stimulation) is used to ensure that their physical properties can be controlled as much as possible.

Video stimuli have been used in EEG/MEG studies involving sign languages (Baus, Gutiérrez & Carreiras 2014; Hosemann et al. 2013), but their adaptation to the techniques' requirements is still challenging. The complexity of the video material (compared to still images) adds extra issues to be controlled. The complexity of videos can, for instance, increase the noise that needs to be avoided to record reliable brain data. As the size of the activity arising from the brain is miniscule (in the order of microvolts [µV in EEG]/femtotesla [fT in MEG]), competing electrical activity coming from the body (such as eye movements or heartbeat) or externally (such as electrical activity from other devices needed in the room during the measurements, see Figure 1) is considered noise. In videos, variations such as changes in luminance or color contrasts, can involve additional brain areas and processes to the ones that are intended to be studied. Brain activity is also considered internal noise. Preferably, any kind of noise should be reduced as much as possible to obtain high-quality data.

More problems arise with videos showing human actions, as in the case of signing. Specifically, for SL research, it is challenging to mark very specifically the phases in the production of signs, such as the offset of the preparation phase, the recognition point, or the onset of meaningful parts of signs (see Table 1). This occurs because the borders of these phases are gradient (i.e. not categorical) and may happen at different times across the videos (Jantunen 2015). ERPs can be locked (i.e. onset/zero-point of ERPs) based on these sign production phases (see Table 1) to identify their related brain processing (Emmorey, Midgley & Holcomb 2022). In addition to the difficulties in identifying signs' parts, the exact same duration in all the videos (and signs) cannot be ensured, which can also elicit changes in brain processing (Skukies & Ehinger 2021) and internal noise.

All signs include semantic information during the preparation phase (for a comprehensive review, see Jantunen 2015). This is because the production of the formational components of the sign must begin before the onset of the stroke, the most meaningful phase in sign production. In particular, the parameters of handshape, orientation, and nonmanual elements provide information that reveals the meaning of the whole sign before the stroke. This happens both with isolated signs as well as in connected signing.

Specifically, when planning/creating the stimuli, we identified two main challenges that needed

to be solved before the actual measurements could start. First, the higher complexity of the video material might increase the physiological noise in the EEG signals. Second, the differences in duration between the videos could create large variability within and across conditions beyond the ones that are intended to be studied. If these problems remain unresolved, they can induce effects that can be camouflaged within the studied signal and lead to misguided results and conclusions.

<table>
<tr><th>Home-Excursion</th><th colspan="2">Movement phase</th></tr>
<tr><td>Home position</td><td colspan="2">-</td></tr>
<tr><td rowspan="5">Excursion</td><td colspan="2">Liberation</td></tr>
<tr><td colspan="2">Preparation</td></tr>
<tr><td rowspan="3">The most expressive phase of the sign</td><td>Pre-stroke hold</td></tr>
<tr><td>Stroke</td></tr>
<tr><td>Post-stroke hold</td></tr>
<tr><td colspan="2">Recovery</td></tr>
<tr><td rowspan="2">Home position</td><td colspan="2">Settling</td></tr>
<tr><td colspan="2">-</td></tr>
</table>

Table 1. Motion (excursion) and movement phases of an individual, isolated sign/gesture. All phases are optional, but signs/gestures typically have Preparation, Stroke and Recovery. The table has been adapted from Arendsen, van Doorn & de Ridder (2007: 317). For visualizations of the phases, please see Jantunen (2015).

3.1. Describing the Procedure

To test how SL is processed at the neural level, EEG measurements need to be taken while participants watch signing situations at a natural pace. As mentioned before, videos need to be carefully created to fulfil several conditions arising from the fields of linguistics and cognitive neuroscience (e.g., visual complexity, luminance, length), while still ensuring the naturalistic nature of the communication process. In the next subsections, we explain the procedure in each step (design, recording, and processing) of preparing the video material that could be used as stimuli in a study on SL processing.

3.1. Stimulus Design, Recording, and Processing

As mentioned before, some physical properties of the videos can create significant processing differences at the brain level that are unrelated to the conditions studied. For this reason, luminance conditions and colors should be kept constant across the videos. Additionally, at a psychological level, other characteristics of the videos could influence the EEG results. To avoid this, we recommend including only one signer, wearing the same clothes, and signing in front of the same background across the videos. The signer should stand in a fixed location so that, at the beginning and end of each stimulus, the location of the signer on the screen is the same. In addition, we do not recommend using a black background because it may cause reflection on the viewer's monitor while doing a measurement. This last recommendation is based on the participants' feedback as a black background was used in our study. We cannot say that the black background would affect the EEG measurements, but it might affect the participant's well-being during the measurements. In our opinion, a light grey background color would work better for this purpose.

Other SL-related characteristics included in the video should be carefully planned. For example, the signing rate (signs per second) should be natural and approximately similar among different stimuli. Articulation should be controlled when it comes to the use of, for example, different facial expressions, mouth actions, and body movements. In addition, the signer's hands and body should be

brought back to an identical neutral position after the signing of each stimulus and left there for at least three to five seconds before continuing the signing. In this way, later in the editing phase, this part can be shortened according to the requirements of the study. Not many more characteristics of the stimuli need to be controlled if they comprise single signs, as no other signs are included in the conditions of the study.

However, if the video stimuli include signed sentences that are, for example, repeated under different conditions, the video materials should be edited so that they are as similar as possible. For example, signers can be asked to watch sentences that are either correct (Condition 1) or semantically incorrect (Condition 2). Then, the brain responses to each kind of sentence are compared. For this purpose, the only thing that should vary between the conditions (kinds of sentences) is the object of the study. In other words, the sentences should be edited so that the same material is used for the unchanging parts of the sentences (i.e., the *sentence frame*), while only the target varies. In addition, when recording varying targets, the signer should produce them in the whole sentence. In this way, the target sign's articulation will be natural in its preparation and offset phases so that it can be edited (if needed) to the sentence frame with a naturalistic result.

Editing should also be considered when choosing the signs for the sentences. For example, large differences in the vertical locations of the signs (hands) before and after the editing point should be avoided. In addition, signs produced before the target should not include a highly repetitive or otherwise varying movement trajectory or a less fixed vertical location in the signing space (e.g., signs CAR, SEARCH, PEN in FinSL; Finnish Signbank, the University of Jyväskylä, Sign Language Center 2018). Furthermore, the aperture of the eyes should be controlled so that it does not vary drastically before and after the editing point. These characteristics can cause a stronger glitch at the editing point of the stimulus. Instead, signs that end with a clear hold (with or without contact with the signer's body) and/or signs that are produced in a relatively fixed location on the signer's body or the signing space are suitable options to be used prior to the editing point (e.g., signs GRANDMA, WOMAN, SHOES in FinSL). Using these types of signs ensures that the position of the signer's hand(s) is as similar as possible between the sentence frame and the utterance from which the target sign is added to the frame. This ensures that the editing of the video results in as natural a result as possible. Finally, if the stimuli also include sentences that do not need editing, an artificial glitch (a point that visually seems like an editing point) should be added to those to avoid variation between the conditions that can affect the results.

All in all, the signer needs to pay attention to many aspects of the contents and articulation of signing while producing the materials, so visual aids – such as a prompter or a large screen – should be used in the recording situation to maximize the control of the variation in the signing. As expressed before, when SL video material is going to be used for research in the linguistic field, the video start and end points need to show how the hands rise from or return to the initial position. Typically, to get a naturalistic result, videos are edited so that they include at least 15–20 frames before and after the production of the sign/sentence. For cognitive neuroscience, that time is significantly longer, based on the high temporal resolution of the functional techniques (EEG and MEG). When creating SL-related videos to be used in cognitive neuroscience studies, our recommendation is to reduce that time as much as possible while keeping the initial position of the signer at the beginning and the end of the videos. In this way, the beginning of the video closely represents the beginning of the sign. We have found that the production of the sign can start as early as three frames after the start of the video while not having a negative effect on the comprehension of the signs.

3.2. Temporal Alignment of Video and EEG Data

After the video stimuli are recorded and edited, they should be synchronized with the EEG/MEG data. The timing properties of the videos are crucial to successfully performing this task. Depending on the research questions and analysis methods, not only could the beginning of the videos be important, but also other components of the signs, such as signing onset and offset times, and duration.

However, what happens if the phenomenon studied includes duration differences? As this was our case while studying semantic processing at the brain level, we tried to ensure that while the videos showed differences in duration across conditions, the stroke (the expressive part of the sign, according to Kita, Gijn & Hulst 1998) onset across conditions was not significantly different. Thus, the beginning of the stroke marks the "definite beginning of meaning." However, in general, the meaning is already comprehensible during the preparation phase preceding the stroke (Jantunen 2015). This has also been confirmed by some cognitive neuroscience studies looking at prediction in signed sentences (Hosemann et al. 2013). If the beginning of meaning does not differ between conditions, even though they have differing durations, their semantic processing at the brain level can be studied with stimulus-locked event-related potentials (ERPs; for more information, see Hernández, Puupponen & Jantunen 2022). This is because the semantic processing at the brain level should be completed around 500–550 ms after the beginning of meaning (around 700–800 ms after the video onset) at the sign level. This time is fairly before the whole sign is performed (around a couple of seconds after the video onset), according to previous studies performed in SL research (Emmorey, Midgley & Holcomb 2022; Ortega, Özyürek & Peeters 2020).

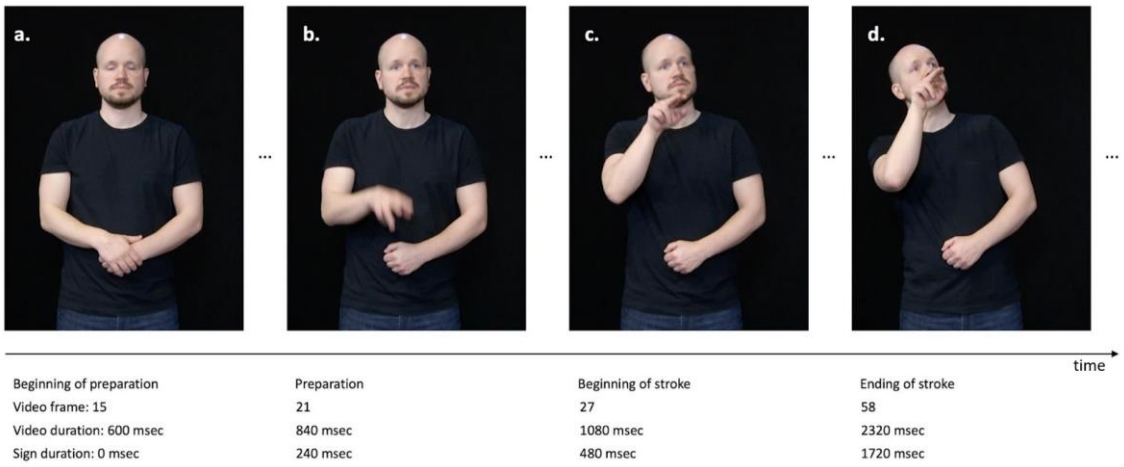


Figure 2. Sign phases (see Table 1) demonstrated with frames (25 fps) extracted from the stimulus video sign LOOK produced with simultaneous enactment

We identified the beginnings of meaning with the help of strokes from the finished video stimulus material. In practice, the beginning of the stroke corresponds to the video frame in which the hands' movement first changes its direction toward the phonemic location of the sign (e.g., Kita, Gijn & Hulst 1998; Jantunen 2015). In Figure 2, this is frame c. From the stroke onset frame, we then moved backward frame by frame toward the moment when the phonemic handshape of the sign was first identifiable. This moment typically occurred with a shift in the eye gaze away from the camera. In Figure 2, this moment of meaning beginning is identified in frame b.

When identifying the sign phases, it is essential to be able to play the video frame by frame and mark the frame number from the video onset. After the frame one is looking for has been identified

(e.g., frame number 27), it can be converted to ms by multiplying the frame number with the 1,000 ms per used frame per second (fps) ratio (e.g., 27 frames x 1,000 ms/25 fps = 1,080 ms, assuming that EEG equipment and camera clocks are synchronized).

4. Combining SL Linguistics and Cognitive Neuroscience in a Pilot Measurement

Previously described measures were implemented in an oddball paradigm (for more details about this kind of paradigm, see Hernández, Puupponen & Jantunen 2022) and tested with an L1 signer pilot participant. In the task, videos of signs (ranging in duration from 1.5 to 2 seconds) were serially presented in random order. Thirty videos showing lexical signs were frequently (82% of probability) presented, while the other 30 videos showing lexical signs with enactment were rarely (18%) presented, with an interval of 500 ms between them. For signs without enactment, the meaning onset started at 320 ms, while for signs with enactment, the meaning onset started at 360 ms. No significant differences were found between the meaning onset times of either condition (frequent and infrequent signs). The participants were asked to press a button as soon as they noticed that "something else" was added to the signs. The resulting EEG responses are shown in Figure 3. The expected ERP (P3b) was visible in the waveforms for the infrequent signs peaking around 710 ms from the video onset and around 350 ms from the beginning of the meaning onset, thus showing its typical characteristics (Hernández, Puupponen & Jantunen 2022).

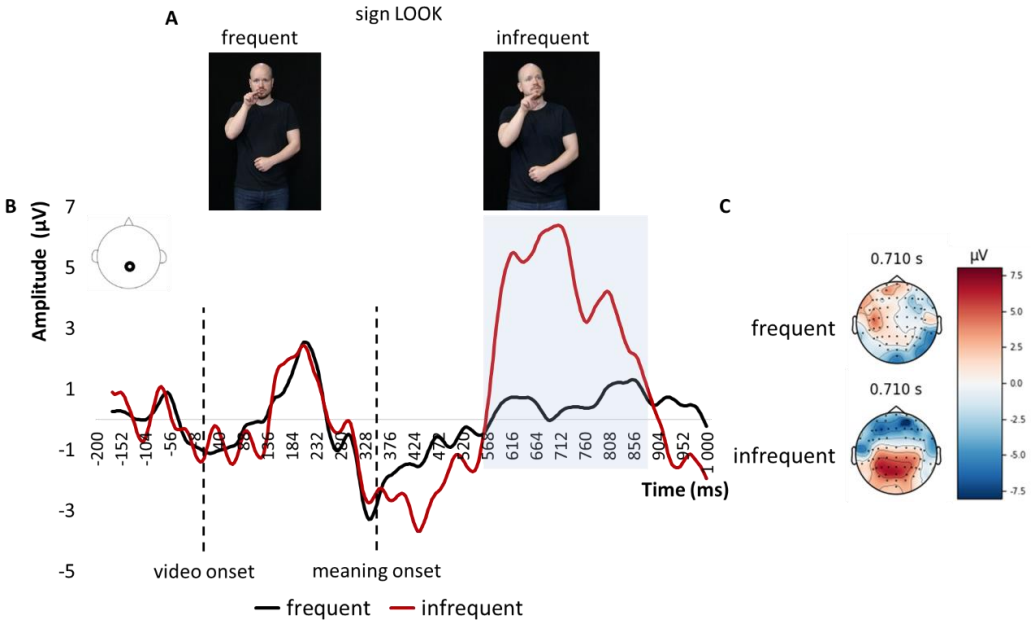


Figure 3. Brain responses from a pilot participant to the described task. A) Example of the kinds of signs used as frequent and infrequent. B) Average waveforms for frequent (black) and infrequent (red) signs in the centroparietal electrode (Pz). The amplitude of the waveforms (in μV) is shown on the y axis, while the latency (in ms) is shown on the x axis. P3b is highlighted in grey. The dashed lines represent the video onset at 0 ms and the meaning onset at 360 ms from the video onset. C) Topographic distribution in the scalp of P3b around 710 ms for the frequent and infrequent signs.

While applying the suggested procedure described in the current paper, we were able to effectively record the intended ERP (P3b). We interpret this as proof that the procedure is successful.

5. Conclusion

This paper has shown our experiences while preparing the stimulus material for the EEG substudy of the ShowTell project. The key challenges (video complexity and duration variability) that we encountered during this research stage were described according to each stage (planning, recording, and processing) of the video stimulus creation process. The solutions given for each of these problems have been described and proved to be successful for our study. Therefore, we are not claiming that our solutions would be the only or the best ones. We hope these will be reviewed and developed in future studies.

6. Acknowledgements

Funding from the Academy of Finland under Project 339268 (ShowTell) is gratefully acknowledged. The authors wish to thank Josh Seligman for the English Language checking of this manuscript.

References

Arendsen, J., van Doorn, A. J., & de Ridder, H. (2007). When and how well do people see the onset of gestures? *Gesture* 7(3), 305-342.

Baus, C., Gutiérrez, E., & Carreiras, M. (2014). The role of syllables in sign language production. *Frontiers in Psychology* 5, 1254.

Emmorey, K., Midgley, K. J., & Holcomb, P. J. (2022). Tracking the time course of sign recognition using ERP repetition priming. *Psychophysiology* 59 (3) e13975.

Hernández, D., Puupponen, A., & Jantunen, T. (2022). The contribution of event-related potentials to the understanding of sign language processing and production in the brain: Experimental evidence and future directions. *Frontiers in Communication* 40.

Hosemann, J., Herrmann, A., Steinbach, M., Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2013). Lexical prediction via forward models: N400 evidence from German Sign Language. *Neuropsychologia* 51 (11), 2224–2237.

Jantunen, T. (2015). How long is the sign? *Linguistics* 53 (1), 93–124.

Kita, S., Gijn, I. V., & Hulst, H. V. D. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In International Gesture Workshop. Springer, Berlin, Heidelberg, 23–35.

Luck, S. J. (2014). An introduction to the event-related potential technique. MIT Press.

Orfanidou, E., Woll, B., & Morgan, G. (2015). *Research Methods in Sign Language Studies*. West Sussex: John Wiley & Sons, Inc.

Ortega, G., Özyürek, A., & Peeters, D. (2020). Iconic gestures serve as manual cognates in hearing second language learners of a sign language: An ERP study. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 46 (3), 403.

Salonen, J., Kronqvist, A., & Jantunen, T. (2020). The corpus of Finnish Sign Language. In Efthimiou, E., F. Stavroula-Evita, T. Hanke, J. Hochgesang, J. Kristoffersen, & M. Mesch (eds.), *Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*. Paris: European Language Resources Association (ELRA), 197–202.

Skukies, R., & Ehinger, B. (2021). Modelling event duration and overlap during EEG analysis. *Journal of Vision* 21 (9), 2037–2037.

The University of Jyväskylä, Sign Language Centre (2018). Finnish Signbank. Available in the Language Bank of Finland (Kielipankki). https://signbank.csc.fi. [Last accessed 5.3.2022].

Towards a Transcription of Modes, Bodies, and Sounds: A Multimodal Actor-Network Theory Informed Transcription of Twitch Digital Discourse

Sarah C. Jackson

The Pennsylvania State University, Department of Applied Linguistics

E-mail: scj5074@psu.edu

Abstract

In the past decade, researchers in both ludology and linguistics have focused on the inclusion of multiple discursive modes, embodiment, as well as non-human elements in the analysis of digital interaction. Digital data sources like Twitch.tv have the potential to give researchers insights into interactions online that span many semiotic modes and concurrent discourses. This platform allows communication through multiple modalities: chat, Voice over Internet Protocol, and elements of the game environment. However, no studies to date have successfully integrated all modes into analysis and proposed a transcription method to represent them. This problem requires the development of a detailed, theory-informed transcription methodology that encapsulates the complex interactions in these multimodal virtual ecologies (virtual worlds) and the communications among players that traverse them. The current paper proposes an approach to the process of transcription in the analysis of interactions as they manifest in Massively Multiplayer Online Role-Playing Games livestreamed on the platform Twitch.tv. This reflects the growing interest in the developing field of "ludolinguistics." Taking a transdisciplinary approach, the proposed transcription methodology employs the basic tenets of Actor-Network Theory. It is also influenced by notions of multimodality, which are enhanced for Twitch.tv, and draws on classic transcription methods and approaches. Transcription exemplars taken from ongoing research into multimodal livestreaming communication are used to clarify and support choices made in the construction of the table-based schema. Finally, implications and practical usages are posited.

Keywords: transcription, multimodality, actor-networks, digital discourse, livestreaming, ludolinguistics

1. Introduction

This project is part of an ongoing academic exploration of the potential that digital discourses have in the study of multimodal communication and interaction. There are many proposed methods of transcribing multimodal phenomena for representation and analyses. However, none adequately addresses the multiple discursive modes and interlocutors that construct and partake in digitally mediated interactions. To the end, I created a transcription method that is systematic in approach and one that can be used to foreground data representation and analysis of multimodal digital communication.

The proposed method draws from classic transcription approaches to multimodal discourses (Ochs 1979; Jefferson 1984; Mondada 2001; Jefferson 2004; Baldry & Thibault 2006), transcription ideologies (Du Bois 1991; Bucholtz 2000), communicative practices across modalities (Goodwin 1994; Kress & van Leeuwen 2001; 2006; Kress 2010; 2012; Gee 2015; Bateman 2021), and recent work on gaming transcriptions (Mondada 2012; 2013; Piirainen-Marsh & Tainio 2014; Baldauf-Quilliatre & de Carvajal 2015; Recktenwald 2017; Rusk & Ståhl 2020). These other transcription methods have included multimodal stills of players and the game. However, they are not necessarily sequentially preserved in data representations, meaning they are usually included outside of the transcription method or not in chronological order with the other transcription elements (e.g., Mondada 2012). On the other hand, some have argued stills are not necessary and include only spoken and written discourses with ad hoc game action descriptions (e.g., Recktenwald 2017). The proposed method, however, differs from prior ones in that it attempts to bring a variety of methods together in an integrated approach. While linguistic and non-linguistic modalities receive attention, the proposed method also incorporates visuals within the sequence of the transcription. It allows for the inclusion of human and non-human interlocutors as well as multiple languages. Ultimately, this method

attempts to give equal consideration to all discursive modes and agents, creating a method that reflects both interests of ludology and linguistics to better address recent attention to "ludolinguistic" approaches (Ensslin 2012; Heritage 2020).

Specifically, the transcription procedure suggested here makes use of intricate data from Massively Multiplayer Online Role-Playing Games (MMORPGs) streamed on the platform *Twitch.tv*. Drawing on elements of Actor-Network Theory (ANT) and multimodality, the proposed transcription method proffers a detailed, narrow, yet highly adaptable transcription of digital interaction. I argue that this method can be utilized outside of these contexts in a multitude of digital settings and discourses. The *Twitch.tv*-MMORPG context represents some of the most complex multimodal technologically mediated communication currently widely accessible on the internet. Therefore, it is used as the focus and demonstrates the ability of this transcription method to capture even some of the most complicated multimodal interactions.

This paper will begin with background information about gaming interactions, MMORPGs, and a fundamental description of the *Twitch.tv* streaming platform. Next, it will explore cornerstone research on multimodal transcription methods that have influenced the current methodology. Then, an exploration of how meaning is made and maintained across digital modes is undertaken, drawing on perspectives of ANT and multimodality. The proposed method will then be explained in detail. Finally, implications and applications of the new method are proposed.

2. Background

Gaming has taken many forms, from the earliest card and board/tabletop games (e.g., checkers) to early videogaming (e.g., *Pacman* 1980) to more common videogame play today, with global audiences (e.g., *Twitch.tv*, eSports) and advanced multiplayer online games (e.g., *World of Warcraft* 2004). Gaming is a distinctive interaction. It can involve a single human player against the game (e.g., *Fallout: New Vegas* 2010), as well as against at least one other human player. Broadly, games can be defined as the contest to achieve a specific goal within the confines of the basic components of the game (e.g., board, cards, moving pieces, virtual worlds, avatars) and its rules (Huizinga 1949; Suits 1967; Rowe 1992; see Stenros 2017). Games have long held as the locus of unique interactional and communicative spaces. Though the gaming modes and audiences differ, each fills the desires of winning, competing, and entertainment (Vorderer, Hartmann & Klimmt 2003; Zagal, Debus & Cardona-Rivera 2019). How each interaction unfolds and via which discursive modes is contingent upon the game and what role the spectatorship takes, if any. Games, in short, create unique interactional environments ripe for further academic study, with implications that reach further than the edges of the gameboard or the limits of digital gaming world.

2.1. MMORPGs

Massively Multiplayer Online games (MMOs; also, MMOGs) are a large subgenre of videogame. These games offer complex multiplayer components that often necessitate collaborative gameplay among players who are not physically co-present. They are a part of a long history of alternate worlds and brought into popular imagination through science fiction/fantasy literature (e.g., *The Lord of the Rings,* Tolkien 1954–1955) and tabletop gaming like *Dungeons and Dragons* (Gygax & Arneson 1974). MMOs stem directly from the creation of Multi-User Dungeons (MUDs). Essex University students created the first MUD in 1979. This was the first time that multiple players could be "together" in the same gaming space despite not being physically co-present. Later technological improvements

gave rise to what is now known as the MMORPG. The first MMORPG was created 1991: *Neverwinter Nights*. The development of the first MMORPG brought with it new and exciting forms of gaming interaction. These games now included graphical representations of the gaming world, along with an avatar, statistical information, gaming interface, overarching narrative, and text-based chat, adding new facets to gaming enjoyment and culture (Vorderer Hartmann & Klimmt 2003; Klimmt & Hartmann 2008; Keating & Sunakawa 2010). Never had so many different elements been brought together in a single gaming environment. Then, in the early 2000s the groundbreaking MMORPG *World of Warcraft* was released by Blizzard Entertainment (2004). Players from all over the world come together in MMORPGs. In these environments, interaction becomes particularly important. Players must maneuver through and successfully utilize all elements at their disposal effectively for favorable outcomes in the game. Groups that come together to complete complicated in-game feats or gamers that become professional streamers typically consist of highly motivated, extremely experienced players.

2.2. Livestreaming and Twitch.tv

Videogame livestreaming is a popular form of online entertainment that involves the player of a game broadcasting their in-game exploits to a live virtual audience. One platform primarily used for this purpose is *Twitch.tv*. This platform, usually referred to as simply *Twitch*, is an "interactive livestreaming service for content spanning gaming, entertainment, sports, music, and more" (Twitch Interactive, Inc. 2022). As both a quasi-social media and livestreaming platform, it has several unique elements. This makes *Twitch* a complex and multifaceted interactional space. It has characteristics of gaming, live entertainment, social interaction, and casual conversation. It is also a form of Computer Mediated Communication (CMC) (Herring 1996, 1). Typically, the basic components of a *Twitch* stream include the streamer's webcam view, the streamer's perspective of the gameplay itself or screen, and the chat from viewers. Figure 1., adapted from Recktenwald (2017), displays the typical view of a *Twitch* livestream which includes the modes of communication facilitated by the platform.

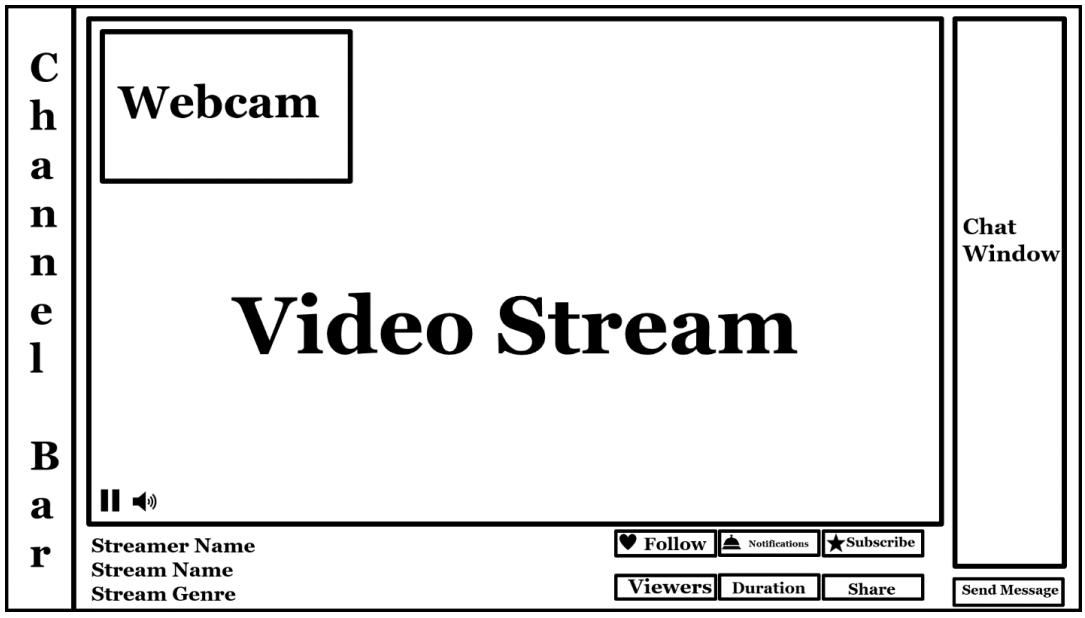


Figure 1: *Twitch* Livestream Layout Schema. Adapted from Recktenwald (2017).

A 'mode,' or 'modality,' broadly speaking, is a communicative act or artifact that carries meaning (Kress & van Leeuwen 2006). Most acts of communication are carried out through an ensemble of

'modes,' wherein each mode plays a part in the creation and maintenance of meaning in a given interaction to complete representational/ communicative tasks (Kress 2010, 28). Commonly used instances of modes include-speech, gestures, music, text, color, etc. This 'ensemble' of modes is made particularly evident in *Twitch* interactions. As the streamer plays a game, for example, audience members can interact through the affordances of the platform via a text-based messaging system. Often, especially in the case of MMORPGs, non-streamer co-players are also added to the mix via Voice over Internet Protocol (VoIP) software like *Discord*, accessed through hardware devices such as gaming headsets with input/output. Therefore, *Twitch* is a highly complicated, multilayered, multi-agent, and multimodal digital discursive space. Table 1 displays these modes in detail, outlining their primary components and material properties.

Mode	**Description**	**Material Properties**
Chat	• accessible by the streamer and audience • used for communication between audience members and streamer as well as among audience members • typically takes the form of orthographic communication via text and emoji	Static
VoIP	• spoken communication among co-player(s) and streamer • Shifts quickly and unfolds temporally as relevance to game-based goals becomes the priority	Temporal
Webcam	• camera view of embodied action from streamer(s) • typically, from the torso up	Temporal
Game Text and Notifications	• non-diegetic texts (Galloway 2006), such as subtitles, mod announcements, health bars etc. • text notifications from *Twitch*, such as new subscribers etc.	Static
Game	• in-game interactions between avatars, NPCs (Nonplayer Characters), the virtual ecology (game environment), strategic markers, maps, objects etc. • contains elements of both static/ stable communication and temporally unfolding/ ephemeral properties (Bateman 2021)	Static/ Temporal

Table 1: Modes and Material Properties of *Twitch* Communicative Affordances

Bateman (2021) suggests that the distinction between 'digital' and 'non-digital' properties of material semiotics is not always clearly defined nor are they straightforward in their composition. Discourses in these technologically mediated spaces are never stable and events evolve rapidly in their both temporal and causal relationships (Baldry & Thibault 2006). However, some of the modes stay rather static, i.e., the mode remains more constant in its communicative goals or stays longer in the communicative space and are interspersed with the more ephemeral and temporally unfolding modes (Juul 2004), i.e., the mode is less static and ever-present but is occasionally the locus of meaning shifts, topic shifts, and dictate subjects of discussion in linguistic mode. The modes present

in *Twitch* discourses, draw attention to the assertion that digital vs. non-digital should not be viewed as a distinct dichotomy, but rather as a spectrum. Some modes take place exclusively in what might be considered the 'digital' realm (e.g., chat, gaming interactions) and some are rather hybridized (e.g., streamer discourse, co-player discourse, streamer responses to chat) (Bateman 2021). All modes and agents (spanning the spectrum of human and non-human) contribute to the creation and maintenance of meaning as they unfold through time and (digital) space.

3. Previous Methods: Transcribing Multimodal Interactions

The cornerstone piece on transcribing multimodal interaction is Ochs' (1979) table design. This brought into interactional analysis the consideration and representation of non-verbal communicative modes, which had previously been largely ignored (Ochs 1979, 172). Ochs' design gave each interlocutor in the interaction a column. These columns were sub-divided into non-verbal and verbal columns. It also introduced transcription conventions in the form of symbols. The symbols notated paralinguistic features, such as intonation (prosody), eye gaze, and embodied movement. This simple design laid the groundwork for the subsequent development of multimodal transcription methods and invited scholars to continuously adapt the method to meet the needs of new and emerging research concerns and technologies (Ochs 1979, 176).

Jefferson (1984; 2004) from a conversational analytic (CA) perspective, expanded the symbols used to represent the intricacies underlying the organization of turn-taking which necessarily involved a narrow focus on paralinguistic features. This established the widely utilized "Jeffersonian Transcription Conventions," which became accepted as the typical approach to talk-in-interaction (Bucholtz 2000). A further expansion was to the use of visual stills/ photos of interactions. Goodwin (1994) showed the importance of artifacts (e.g., tool use), manipulation of the material environment, and embodied practice was vital in the representation and analysis of interaction. Critically, he used photos in his analysis, usually not integrated into the transcription itself but included nearby. With the addition of these different modes of interaction, Goodwin concluded that these elements are essential in understanding human interaction and cognition as "socially situated phenomena" that contribute to the creation and maintenance of meaning (1994, 471).

Baldry and Thibault (2006) made copious contributions to the areas of multimodal transcription taking inspiration from Ochs' tabular format. These explorations of multimodal data representations through transcription and subsequent text analysis drew attention to the multilayered construction of meaning in communicative multimedia artifacts such as webpages, films, and written texts. They sought to bring out the micro- and macro-level meaning making processes across modes with their methods that covered a myriad of data types and expanding technologies. Each column represented a different type of temporally co-occurring communicative phenomenon such as visual stills and their descriptions, kinesic (embodied) action, and soundtrack/verbal discourse. Critically, visuals were kept in sequence with other semiotic modes. Nevertheless, the ever forward march of technology since the inception of these methods rendered some aspects slightly unfitting for many types of more recent developments in gaming, social media, and other forms of CMC.

A more detailed multimodal transcription, which attempted to maintain sequence of interaction, description of embodied action, paralinguistic features, as well as transcript-integrated stills/ photos was developed by Mondada (2001). Though stills/photos are not always maintained in the sequence of the transcript but placed elsewhere, Mondada made great strides in representing interactional data across multiple modalities and among human *and* non-human agents. Mondada was also one of the first to bring in an integrated approach to video gaming interactions among co-present players (2012;

2013). This method was mirrored in subsequent studies, which concentrated on multimodal player interactions and incorporated non-human agents (Piirainen-Marsh & Tainio 2014; Baldauf-Quilliatre & de Carvajal 2015; Rusk & Ståhl 2020). These studies brought in elements of the gaming world, or virtual ecology, into analysis. However, players in these studies were not engaged in technologically mediated communication.

Recktenwald (2017) laid critical groundwork for addressing this gap in the study of digitally mediated discourses and interactions. He developed a table-based transcription method, harkening back to the design of Ochs (1979). This broad method enabled transcription of *Twitch* discourses and interactions. He utilized separate columns for separate discursive modes, including "Game Events," Streamer discourse and action ("Streamer"), and text/emoji commentary from the audience ("Chat"). The sequence of these events was maintained with the "Timestamp" column. In most basic terms, Recktenwald (2017) was able to capture the primary elements of complicated *Twitch* interactions and made great improvements in describing interactions as they unfold in this discursive space. But Recktenwald did not include visuals and did not address *Twitch*-sourced data that involved multiple co-players engaged in multimodal communication via the instrumentality of VoIP in addition to the game environment.

This leaves a considerable gap in the micro-level analysis of these already rather complicated technologically mediated discourses. Co-player interaction adds another layer to the meaning making process and is essential in understanding a vast array of mediated discourses. And visuals, though addressed by some of the previous methods, have not been proposed in multimodal CMC platform transcriptions. Indeed, as previously discussed, *Twitch* has multiple modes of communication among agents, and visuals are included in at least three (the game view, the webcam view, and chat). Additionally, discourses – particularly videogaming-centered discourses – involve the influence of multiple artifacts and non-human agents (see Mondada 2012; Piirainen-Marsh & Tainio 2014). Baldry and Thibault's method, for example, includes visual descriptions of non-human agents in interactions but does not position them as participants in the interaction that have the power to not only communicate, but also to enact change in an interaction. Finally, with the exception of Mondada and a few others' work, multimodal transcription methods do not commonly suggest ways of addressing multilingual/translingual contexts or glosses/translations in technologically mediated contexts. Perhaps quite obviously, gaming and social media discourses span the globe and more commonly involve the use of not only multiple modes of communication, but multiple languages. Despite this, few transcription methods of CMC data have suggested ways of including multiple languages.

The proposed transcription technique seeks to further expand and continue the lineage of the development of multimodal transcription methods. It is fine-grained and narrow in approach and draws from previous transcription schemas. It creates a flexible, highly adaptable procedure for representing and analyzing complex discourses, multimodality, embodiment, technologically mediated discourses and multilingual contexts.

4. Making and Maintaining Meaning Across Modes and Among Agents

In *Twitch*-MMORPG gaming environments, players must deftly navigate complex webs and layers of shifting discourses across multiple modes. These modes must be properly maintained to complete gaming objectives and engage in successful interactions with co-players and audience. Meaning making and maintenance, or cohesion, across these multiple modes is an important phenomenon in digital discourse (Recktenwald 2017; Steinkuehler 2008). Trying to describe and transcribe these phenomena without an orienting theoretical frame to explicate and explore it, however, is difficult.

Intersubjectivity, in basic terms, describes how meaning is maintained in situated interactions (Goodwin 1994; Gee 1999). But the shifting modes and varying interlocutors in these environments require more explanatory and descriptive power. Most theories, for example, do not include non-human agents in interactions nor do they transcribe the discourse they produce. They also often ignore intricate multimodal discursive communicative practices. One theory that has widely been used in videogame studies and can fulfill the needs of this transcription method is Actor-Network Theory (ANT).

ANT, sometimes referred to as material-semiotics (Law 2019), posits that all things, material and conceptual, human and non-human, exist in ephemeral, perpetually reconfiguring networks of relationships and discourses (Callon 1986; Latour 1993; 2005; Law 2019). All parts/factors that make up an interaction or social network are of equal importance: actors involved, objects, humans, ideas etc., are agentive. They are referred to as 'agents' in these communicative interactions. Therefore, it promotes equal consideration of all these elements in an interaction without giving particular attention to one agent or one mode of communication. Discourse and agency must be considered in terms of a spectrum instead of the typical dichotomies of linguistic vs. non-linguistic, human vs. non-human.

To complement ANT, multimodality was chosen in the construction of the transcription schema. As previously discussed, multimodality is an approach to communication that includes the analysis of meaning through multiple meaning-bearing resources across modes; meaning and communication are not limited to linguistic forms but include all meaning-making sources (Kress & van Leeuwen 2001; 2006; Kress 2010; 2012; Gee 2015). This includes images, aural input, embodiment (gaze, gesture, body orientation), etc. It is the analysis of the discourse produced through the synergistic blend of all media accompanying the linguistic elements and how those elements combine to make meaning. Particularly pertinent to the study of *Twitch* videogaming, this orientation calls for the inclusion of important game elements as "discourse" (Ensslin 2012; Gee 2015). Besides the human elements, their avatar (inter)actions, sound effects/music, and textual manifestations of discourse, multimodality also allows for the analysis of colors, codes, symbols (Kress & van Leeuwen 2001; 2006; Kress 2010; 2012), positions, markers, and (virtual) environmental manipulations (see Goodwin 1994).

Multimodality and ANT are used in concert to give structure to the often complex and nebulous constellation of agents and modes that build these interactions. This strategic combination seeks to better capture and represent the components in this plurality of modes, agents, and spaces – to represent data in a way that reflects a spectrum or continuum rather than strict dichotomies between the digital and non-digital, human and non-human, linguistic and non-linguistic. It complexifies the notion of participation framework (Wenger 2011) and pressures the researcher to consider this in analysis. Just as ANT does not ensure that all modes are included in analysis, multimodality does not necessarily ensure the inclusion of multiple agents. Thus, both orientations were used in the construction of this transcription method.

Consequently, ANT is not a theory of analysis, but rather a schematizing surface-level theoretical model. It forces the transcriber to recognize, and therefore represent and analyze, not only multimodal and non-linguistic forms of communication, but non-human communicative agents and the space in which interactions take place. As previously described, livestreamed gaming interactions on *Twitch* are highly complex not only in terms of modalities of communication, but also agents using the modes of communication. Humans, for instance, are not the only communicative agents in these environments (Taylor 2009; Piirainen-Marsh & Tainio 2014; Anderson 2017). Humans, avatars, NPCs, audiences, all contribute to communication and interaction and are thus represented equally in

the transcription. This provides new opportunities to study communication in these interactions, moving beyond representation of just humans and words. ANT and multimodality also help with the flexibility of the method. Depending on the interaction, type of game, physical artifacts used to play or interact, modes of communication etc., components can be easily removed or added to capture the sequence. Viewing modes and agents in this way also contributes to the flexibility of the method, allowing it to potentially avoid obsolescence as technologies develop and change (Bateman 2021).

5. Transcribing Multimodal Digital Interactions

As noted, digital interactions in *Twitch* livestreams are particularly complex. In conjunction with previous transcription methods, ANT, and multimodality facilitated the development of the proposed transcription method explored here. To accommodate different categories of modes and agents within digital interactions, a table-based schema was selected, as inspired by Ochs (1979) and Recktenwald (2017).

In the table-based system, the sequence of the interaction is maintained top to bottom and concurrent modes of discourse are preserved left to right. Given the flexibility of the design, it can be manipulated based on need. For example, Jeffersonian conventions that focus on narrow representations of speech sounds and pauses may not be relevant to a particular study, or chat may not be relevant or available in other interactions. Categories can easily be changed to accommodate different mediated or face-to-face discourses, not just gaming contexts.

In order to fit within the chart and keep the transcription from becoming too chaotic in appearance, visuals should be kept a reasonable size. Commentary and descriptions can be added where needed. Overall, the transcription method seeks to not only capture complex gaming interactions so that they can be analyzed and studied closely with multimodality, but also remain in keeping with the tenets of the proposed foundational theory of ANT. This way, all discursive elements and agents can be given equal consideration and the transcript can show how many of these overlooked elements combine to make and maintain meaning.

5.1. Basic Layout

Figure 2 is a representation of this transcription method and shows the basic proposed configuration. Each column has a heading, under which the different modes of communication are categorized, transcribed, and displayed preserving their sequence.

The area to the far left labeled “Temporality” is the space for timekeeping measures used to show the chronological sequence of interaction. In the case of online streaming, it will typically take the form of a “timestamp.” Actions and discourse can thus be mapped across modes and among different agents. For instance, when an NPC’s (Nonplayer Character) talk overlaps with a player’s, chronological sequence can still be preserved. When using timestamps, the chronology can be maintained by durations (e.g., 04:10-04:15; see Figure 4) or by single seconds, wherein each line is one second (see Figure 2 and Figure 3). Parts of excerpts can also be referenced within the analysis by the temporality durations (e.g., 11:15-17 and at 12:09).

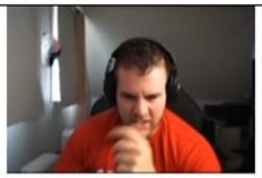
Temporality	Audible Discourse / Visual Text	Embodiment	Game Action
00:01	*PL1: ΔRobinΔ this - this is just way easi[er↘] *PL3: °[that's]	 PL1 oriented towards screen gaze on map, hand on face (00:00-00:006)	 Map mod with player-made markings in various colors is brought up and actively referenced to (00:01- 00:06)
00:02	*PL3: that's xxx→°		
00:03	*PL1: this- this gr- (.) just pull		
00:04	*PL1: four↘ like u:h →		
00:05	*PL1: you're eliminating ∇two∇ ↑caster packs↘		
00:06	*PL1: °↓or whatever ∇this∇ patrol that's [like→]° *PL3:[∇yea b]		
00:07	*PL 3: utΔ	 PL1 moves hand from face at 00:07	 PL1 uses mouse to reference "caster packs" that they need to eliminate, indicated by different colors on the displayed map. PL1 mouses over "Depraved Houndmaster" enemy icon at 00:007

Figure 2: Basic Layout. Data from *World of Warcraft*.

"Audible Discourse" is any relevant sound made by the game, mods, avatars, weapons, players, or technology that can be transcribed using (in this case) the Latin alphabet (onomatopoeia etc.). "Visual Text" is simply any non-diegetic text (non-narrative; text that in-game characters cannot "see") that may appear in the visual field of the virtual ecology, such as an announcement or character subtitles. This does not include text-based chat, which can be accounted for in the next version (see Figure 2). Each relevant aural/visual discourse is transcribed in actual duration. This means that spoken discourse, for example, can span several lines in the "Temporality" column. The distinction between "Embodiment" and "Game Action" is critical. "Embodiment" refers to any corporeal action (gesture, movement, eye gaze) as it manifests in human or non-human, virtual forms. "Game Action," on the other hand, should be considered as any arrangement of multimodal events that transpire involving multiple elements that are relevant for meaning making or analysis.

5.2. Inclusion of Chat

Because the games are livestreamed via *Twitch*, viewers can also participate in commentary as the game is underway. To annotate/record relevant chat interactions by viewers, I have amended the original table and added a column to the right, entitled "Chat." Figure 3 represents this.

Tempor-ality	Audible Discourse/ Visual Text	Embodiment	Game Action	Chat
3:03:46				StarlightNebula_: Gratz on the clear bro
3:03:47	*MAZZ: I'm not taking it			
3:03:48	*MAZZ: there or whatever			
3:03:49	*MAZZ: I don't			
3:03:50	*MAZZ: think			
3:03:51	*MAZZ: like			
3:03:52	*MAZZ: Revenge is takin' one			flabaebae: They're doing it mom
3:03:53	*MAZZ: just to clarify			
3:03:54				KaiCreator707: hell yeah clean!
3:03:55	*MAZZ: you see that			
3:03:56	*MAZZ: open spot Revenge?		 Open spot in avatar circle formation	
3:03:57	*WOOPS: I mean it's gonna die before			JillTime: POGFISH
3:03:58	*WOOPS: it gets to the DPS			
3:03:59	*laughter*	Avatars coordinate final abilities 		
3:04:00				
3:04:01	*WOOPS: one day			
3:04:02	*WOOPS: clear			Akanowo: clean ezclear
3:04:03	*WOOPS: three hour clear even *NOGA: nice		 Boss dies 3:04:03-6	
3:04:04	*NOGA: all			
3:04:05				
3:04:06	*WOOPS: good job			
3:04:07				
3:04:08	*ANUBIS: sii:			
3:04:09	*ANUBIS: :iick *NOGA: woo			
3:04:10				
3:04:11				
3:04:12		 Avatars celebrate by jumping and running around 3:04:10 -16		Arklord_Xeno: GGs

Figure 3: Inclusion of Chat. Data from Final Fantasy XIV.

The "Chat" relates to the "Temporality" column directly, as they are placed at the timestamps in which they appeared live. Because the program *Chatty* reliably retrieves both text and emoji, both modes can be included in representation and analysis.

5.3. Extended Layout

The components can be easily moved and elaborated on, such as adding columns for translations/ glosses ("Gloss/ Translation" and "Chat Translation" columns) as well as chat or written text where required. This is reflected in Figure 4.

Temporality	Audible Discourse / Visual Text	Gloss/ Translation	Embodiment	Game Action	Chat	Chat Translation
00:00- 3:17	Screen: STREAM STARTET BALD	Stream starts soon			WizeBot ✔ LIVE ONLINE under the game World of Warcraft, the uptime begin.	
3:18					DerBroton: lalalalalalalalalalreeenaaa	
3:19					DerBroton: <<3	
3:22					DerBroton:	
3:59-4:00	Larena5: Hallo!	Hi!	Larena5 appears and waves at 4:00 			
4:00-4:01	Larena5: was geht?	what's going on?		 Avatar appears in Oribos, the Eternal City at 4:01 and remains there		
4:01-4:03	Larena5: Der brobrobroBroton!					
4:06	Larena5: Nja?	Well?				
4:07-4:08	Larena5: Was geht ab?	What's up?				
4:10-4:15	Larena5: *sings* can't turn back time xxx					
4:15					DerBroton: nich viel soweit bei ihnen so	Not much so far, how's it with you?

Figure 4: Extended Layout. Data from World of Warcraft. German language representation.

6. Questions Addressed by this Method

The inclusion of multimodality in the transcription of video gaming is crucial to the understanding of the moment-by-moment unfolding of collaborative interactional gaming in the virtual world. Games inherently involve challenges, obstacles, strategic planning, anticipation of future moves by self and others, and re-evaluation of prior moves by self and others. Video games expand the worlds of space of play, players, challenges and obstacles and can provide real-world insights into human cognition and interaction that spans modes, spaces, and agents. In general, this transcription method allows for a robust analysis and more inclusive reading/representation of data. It ensures that the researcher considers all modes, elements, spaces, and agents in potentially contributing to the unfolding of an interaction, the maintenance and creation of meaning, and as unravelling components that may contribute more deeply to the satisfactory investigation of and responses to research questions.

A rigorous and detailed transcription system that encompasses the broad ranges of potential spaces including non-human interactants, avatars, attacks, lights, sounds and many more, will allow future in-depth explorations into how, specifically, players: interact, perceive and react to threats, strategically anticipate future moves, strategically (re)evaluate prior move(s), guide and direct, teach, accept and reject ideas, employ multilingual resources and many more. It can provide new and complex insights as we delve into such questions as combining the macro-, meso-, and micro-instances of interaction and strategic moves toward collaboration and winning. The following are just a few examples of the types of research questions the proposed method could assist in addressing, questions which are currently being investigated using this very same transcription methodology (Jackson, in progress):

1. How do players maintain intersubjectivity and situated meaning while engaged in multifaceted, ephemeral layers of concurrent discourses?
2. How do players utilize the multimodal communicative affordances of the game to

communicate?

3. How does strategy form and how is it ultimately brought into game action?
4. How do co-players communicate with teammates about the game, immediate obstacles, and future obstacles?
5. How are multiple modes as well as multiple named languages navigated and utilized in mediated interactions?
6. In what ways do non-human agents shape and influence interactions?
7. In what ways does the quality of interaction change as the players: lose, try again, win, come close to winning?
8. How does the audience orient to, comment on, and engage with the streamer and the unfolding multimodal action?
9. What can the maintenance and utilization of multiple communicative modes reveal about human cognition and interaction in not only technologically mediated spaces, but face-to-face interactional spaces as well?

7. Conclusions and Implications

This transcription method, drawing from tenets of ANT and multimodality, adds important contextual references. It shows how intersubjectivity is maintained in interactional practices that span concurrent modes of discourse and agents. It offers an adaptable, multimodal method that is still systematic, uniform, and theoretically oriented. It can be made narrower or broader according to need.

The digital frontier is still rather neoteric and underexplored considering its sheer vastness and every-changing, ever-growing landscape. "Ludolinguistics," or the field that combines gaming with linguistic study, is also a rather new area of research (Heritage 2020). The proposed transcription method seeks to broaden the horizons of the study of discourses and interactions that span the continuum of digital – non-digital (Bateman 2021), human – non-human (Latour 2005), and linguistic – non-linguistic (Kress & van Leeuwen 2006; Kress 2010). Rather than being viewed as dichotomies, these phenomena should be viewed as spectrums, within which interactions unfold. The strategic combination of ANT and multimodality ensure that data is represented to reflect such spectrums. It also provides a level of flexibility that allows adaptations to new forms of technology. As digital interactions become more complex and commonplace, it is critical that transcription methods be continuously adapted and expanded upon (Ochs 1979). The rise of increasingly complex forms of multimodal, multimedia communication (e.g., TikTok, *Twitch*, multiplayer Virtual Reality (VR) and Augmented Reality (AR) just to name a few) it is essential that methods of transcription be adapted (Ochs 1979) and are created in a way that ensure that they do not create their own obsolescence (Bateman 2021).

The proposed methodology seeks to systematically and rigorously reveal the various ways in which meaning is constructed and incrementally developed in multimodal situated game action as players work together (and against each other) strategizing to gain points, fend off enemies, avoid disaster, and ultimately win the game. Indeed, in contrast to previous methods, more multimodal and even multilingual meaning making processes in these virtual videogaming contexts can be analyzed through finer-grained, higher definition multimodal linguistic lens, shedding light on areas of strategic communicative interactions that have heretofore been neglected.

References

Anderson, S. L. (2017). The corporeal turn: At the intersection of rhetoric, bodies, and video games. *Review of Communication* 17 (1), 18–36.

Baldauf-Quilliatre, H., & de Carvajal, I. C. (2015). Is the avatar considered as a participant by the players? A conversational analysis of multi-player videogames interactions. *PsychNology Journal* 13(2), 127–147.

Bateman, J. A. (2021). What are digital media? *Discourse, Context & Media* 41 (2). https://doi.org/10.1016/j.dcm.2021.100502

Baldry, A., & Thibault, P. J. (2006). *Multimodal transcription and text analysis: A multimedia toolkit and coursebook*. London: Equinox.

Bucholtz, M.. (2000). The politics of transcription. *Journal of Pragmatics* 32, 1439–1465.

Callon, M. (1986). The Sociology of an Actor-Network: The Case of the Electric Vehicle. In M. Callon, J. Law & A. Rip (eds), *Mapping the Dynamics of Science and Technology*. Macmillan Press, London, 19–34.

Du Bois, J. W. (1991). Transcription design principles for spoken discourse research. *Pragmatics* 1(1), 71–106.

Ensslin, A. (2012). *The language of gaming*. Basingstoke: Palgrave Macmillan.

Galloway, A. R. (2006). *Gaming: Essays on algorithmic culture*. Minneapolis: University of Minnesota Press.

Gee, J. P. (1999). *An introduction to discourse analysis: Theory and method*. London: Routledge.

Gee, J. P. (2015). *Unified discourse analysis: Language, reality, virtual worlds, and video games*. London: Routledge.

Goodwin, C. (1994). Professional Vision. *American Anthropologist* 96 (3), 606–33

Gygax, G. & Arneson, D. (1974). *Dungeons and Dragons* (1st ed.). Lake Geneva, WI: TSR, Inc.

Heritage, F. (2020). Applying corpus linguistics to videogame data: Exploring the representation of gender in videogames at a lexical level. *Game studies*, 20 (3).

Herring, S. C. (ed.) (1996). *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives* (Vol. 39). Amsterdam: John Benjamins Publishing Company.

Huizinga, J. (1949). *Homo Ludens: A Study of the Play-Element in Culture*. Abingdon: Routledge & Kegan Paul.

Jackson, S. (in progress). *Multimodal MMORPG Interaction on the Livestreaming Platform Twitch*.

Jefferson, G. (1984). On the organization of laughter in talk about troubles. In J. Atkinson (ed.), *Structures of Social Action*. (Studies in Emotion and Social Interaction). Cambridge: Cambridge University Press, 346–369. DOI:10.1017/CBO9780511665868.021

Jefferson, G. (2004). Glossary of transcript symbols. In G. H. Lerner, *Conversation analysis: Studies from the first generation*. Amsterdam: John Benjamins Publishing Company, 24–31.

Juul, J. (2004). Introduction to Game Time / Time to Play: An examination of game temporality. In N. Wardrip-Fruin & P. Harrigan (eds.), *First person: New media as story, performance, and game*. Cambridge: MIT Press, 131–142.

Keating, E. & Sunakawa, C. (2010). Participation cues: Coordinating activity and collaboration in complex online gaming worlds. *Language in Society*, 39 (3), 331–356.

Klimmt, C. & Hartmann, T. (2008). Mediated interpersonal communication in multiplayer video games: Implications for entertainment and relationship management. In E. A. Konijn, S. Utz, M. Tanis, & S. B. Barnes (eds.), *Mediated interpersonal communication*. New York: Routledge, 309–330.

Kress G. (2010). *Multimodality: A Social Semiotic Approach to Contemporary Communication*. London: Routledge.

Kress, G. (2012). Multimodal discourse analysis. In J. Gee & M. Handford (eds.), *The Routledge Handbook of Discourse Analysis*. London: Routledge, 35–50.

Kress G. and van Leeuwen, T. (2001). *Multimodal Discourse: The Modes and Media of Contemporary Communication*. London: Arnold.

Kress G. and van Leeuwen T. (2006). *Reading Images: The Grammar of Visual Design*. London, New York: Routledge.

Latour, B. (1993). *We Have Never been Modern.* New York: Harvester Wheatsheaf.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford: Oxford University Press.

Law, J. (2019). *Material semiotics*. The Open University, Milton Keynes; Sámi Allaskuvla, Guovdageaidnu. http://www.heterogeneities.net/publications/Law2019MaterialSemiotics.pdf

Mondada, L. (2001). *Conventions for Multimodal Transcription*. https://www.lorenzamondada.net/multimodal-transcription

Mondada, L. (2012). Coordinating action and talk-in-interaction in and out of video games. In R. Ayaß & C. Gerhardt (eds.), *The appropriation of media in everyday life*. Amsterdam: John Benjamins Publishing Company, 231–270.

Mondada, L. (2013). Coordinating mobile action in real time: The timely organisation of directives in video games. In P. Haddington, L. Mondada & M. Nevile (eds.), *Interaction and mobility: Language and the Body in Motion.* Berlin: De Gruyter, 300–342.
Ochs, E. (1979). Transcription as theory. In E. Ochs & B. B. Schieffelin (eds.), *Developmental Pragmatics*. New York: Academic Press, 43–72.
Piirainen-Marsh, A., & Tainio, L. (2014). Asymmetries of knowledge and epistemic change in social gaming interaction. *The Modern Language Journal* 98 (4), 1022–1038.
Recktenwald, D. (2017). Toward a transcription and analysis of live streaming on Twitch. *Journal of Pragmatics* 115, 68–81.
Rowe, M. W. (1992). The definition of 'game'. *Philosophy* 67 (262), 467–479.
Rusk, F., & Ståhl, M. (2020). A CA perspective on kills and deaths in Counter-Strike: Global Offensive video game play. *Social Interaction*. *Video-Based Studies of Human Sociality* 3 (2). https://doi.org/10.7146/si.v3i2.117066
Stenros, J. (2017). The game definition game: A review. *Games and culture* 12 (6), 499–520.
Steinkuehler, C. A. (2008). Massively multiplayer online video gaming as participation in a discourse. *Mind, culture, and activity* 13 (1), 38–52.
Suits, B. (1967). What is a Game? *Philosophy of science* 34 (2), 148–156.
Taylor, T. L. (2009). The assemblage of play. *Games and culture* 4 (4), 331–339.
Tolkien, J. R. R. (1954–1955; 2012). *The Lord of The Rings: one volume*. Boston: Houghton Mifflin Harcourt.
Twitch Interactive, Inc. (2022). *About*. Twitch.tv. Retrieved January 6, 2022, https://www.twitch.tv/p/en/about/
Vorderer, P., Hartmann, T. & Klimmt, C. (2003). Explaining the enjoyment of playing video games: the role of competition. *Proceedings of the second international conference on Entertainment computing.* Carnegie Mellon University, 1–9.
Wenger, E. (2011). *Communities of practice: A brief introduction*.
Zagal, J. P., Debus, M. S. & Cardona-Rivera, R. E. (2019). On the ultimate goals of games: Winning, finishing, and prolonging. *Proceedings of the 13th International Philosophy of Computer Games Conference* (Vol. 11).

A Qualitative and Quantitative Study on Ancient Latin Texts Concerning the Concept of *aether*: Some Methodological Considerations

Henrik Roschier
University of Jyväskylä
E-mail: yrhrosc@student.jyu.fi

Abstract

My study focuses on the semantic analysis of an ambiguous Latin word, *aether*, which may roughly be translated as 'celestial substance'. The study is carried out by utilising both qualitative and quantitative methods on a digital text corpus containing Latin literature dating from ca. 200 BCE to 200 CE. In this paper, the compilation process and the contents of the corpus are described.

Beside qualitative research (close reading), various computational methods are implemented on the corpus (distant reading). Of the latter, cosine similarity and correspondence analysis are discussed in this paper. I will explicate one way of combining qualitative and quantitative methods by assigning numerical values to a table built into a concordance. The values are based on my interpretation of the relevant text passages, and they can be utilised in quantitative analyses as frequency-based entries in a term-document matrix.

It seems evident that 'sky' is the most common denotation of *aether*. Human interpretation is necessary in semantic observation, but significant improvement may be gained by using quantitative methods in parallel.[52]

Keywords: Latin language, corpus linguistics, computational linguistics, semantics, distributional semantics, conceptual history

1. Introduction

My study focuses on the semantic analysis of an ambiguous Latin word, *aether*, which may roughly be translated as 'celestial substance' or 'heaven', depending on the context (Le Boeuffle 1987; TLL 1:1149,13-1152,48). The objective and approach of this research is two-fold. First, the empirical aspect of the study consists of analysing the semantic definitions of *aether* in various contexts. In addition, I seek to elucidate the word's semantic relations to relevant vocabulary. Second, I aim at providing methodological information about combining traditional qualitative research (close reading) and quantitative and computational linguistic methods (distant reading). This paper will discuss some of the methods used.

The time period under scrutiny is ca. 200 BCE - 200 CE, the era during which the bulk of Latin literature was written. It also marks a period of intense cultural relations with the Greek-speaking world, as well as rich intertextuality in written language. For example, *aether* itself is a Greek loanword (from αἰθήρ). In accordance with the geocentric world view prevailing in antiquity, the term *aether* was used both to denote the sky in general, and to describe the region of the stars and planets as opposed to the Earth (Le Boeuffle 1987).

Choosing one particular lexeme as the primary subject of study is an attempt to link semantic analyses to historical context, and to utilise quantitative methods for gaining a better understanding of ancient texts, maybe even of their writers. The properties of the text corpus analysed in this study will be described below.

[52] I would like to thank Hanna-Mari Kupari, Timo Korkiakangas, and two anonymous reviewers for their fruitful comments and suggestions. I also thank Charles Dickins for proofreading.

2. Data and Methods

2.1. The Corpus

2.1.1. Selection and Building

With Latin, the scarcity of surviving texts significantly limits the scope of corpus research. However, the situation is improving, and Latin is relatively well-resourced among ancient languages (McGillivray 2013). Nevertheless, the researcher is in a weaker position compared to modern language studies, with little possibility of collecting massive corpora at will.

Since the goal here is to focus on the usage and meanings of a relatively rare word, it is desirable to gather every occurrence possible, notwithstanding text form or genre. As *aether* is known in advance to be related to the sky and heavenly bodies, I chose a couple of auxiliary words as criteria for text selection. The aim of this was to help in building a detailed semantic profile of *aether* with support from related words, even using texts in which *aether* does not occur.

The primary criterion for selection was that the text had to contain one or more of the following seven lexemes: *aether* 'sky, celestial substance' (460 occurrences in the corpus), *aetherius* 'of celestial substance, heavenly' (173), *aethra* (largely synonymous to *aether* but a rarer word, 34), *astrum* 'star' (616), *sidus* 'star, constellation' (1369), *stella* 'star' (796), and *luna* 'moon' (820). One can easily see that these words include three cognates from the root *aeth(e)r-*. At the same time, the star words build a link to natural philosophy and cosmology, i.e., the world view of the ancient writers. Namely, since Aristotle (384-322 BCE), αἰθήρ / *aether* was often considered to be the exclusive substance of the heavenly bodies, as opposed to the traditional four elements (earth, water, air, fire) composing the Earthly world (Couprie 2011; Lloyd 1970; Wright 1995). Finally, I restricted the vocabulary representing actual celestial bodies to *luna* alone, bearing in mind that ancient celestial nomenclature involves wide-ranging mythological connotations, which could make semantic analyses complicated. For example, the names of planets are also the names of primary Roman divinities (Zucker 2016). Thus, by studying the semantic profile of the selected lexemes, the objective is, beside linguistic purposes, to elucidate the conceptions of celestial bodies featuring among Roman authors.

Chronologically, I restricted the text selection to Apuleius (ca. 123-175 CE) at the latest. Since the third century onwards, the rise of Christian literature produced even wider connotations for words referring to heaven. For now, such diversification was considered undesirable for semantic analyses. However, it would later be a natural extension for this project to trace semantic changes over a longer time. It should be noted, however, that some of the texts are of dubious origin or by unknown authors, so the time period can only be defined approximately. As for the earliest texts, *aether* is first attested in fragments by the poet Ennius (ca. 239-169 BCE). For statistical purposes, the texts were divided into five chronological categories by centuries but separating the prolific period around the beginning of Common Era into its own class (Table 1).

	3rd – 2nd cent. BCE	1st cent. BCE	ca. 30 BCE – 30 CE	1st cent.	2nd cent.
words (unedited)	110 031	1 197 789	817 784	1 733 471	601 533
%	2.5 %	26.9 %	18.3 %	38.9 %	13.5 %
text files	16	74	37	109	19
%	6.3 %	29.0 %	14.5 %	42.7 %	7.5 %

Table 1: Chronological distribution of the texts: the number of texts dating from each period, with their word count.

The chronological distribution is uneven, as the numbers indicate. At a later stage, I will reconsider the categorisation to minimise the loss of information which is an inevitable result of classifying. As there was significant variation in the length of texts, too, some of the longest ones were divided into several text files. The range was still left broad (from 30 to 172 000 words in a single text file).

The texts were also classified into 13 categories by their content, or roughly, topic. Naturally, some of the works do not easily fit into any singular category, as classification is a matter open to deliberation. The analyses presented in this paper do not yet involve the topic categories, but they are listed here so that the reader can see what types of texts the corpus contains (Table 2).

	TOPIC	TITLES	TOKENS	TOKENS %
	agriculture, horticulture	4	203 169	5 %
	astronomy, astrology, meteorology	7	114 436	3 %
	drama, plays	21	148 929	3 %
	epic poetry	12	325 684	7 %
	epistles, literature in letter form	5	342 282	8 %
	history	11	1 005 582	23 %
	mythology	5	205 270	5 %
	nature	5	320 594	7 %
	philosophy	26	567 454	13 %
	poems (other than epic)	29	329 239	7 %
	rhetoric, language, speeches	12	499 602	11 %
	technology, arts, disciplines	5	235 447	5 %
	various, miscellaneous	7	162 920	4 %
SUM	**13**	**149**	**4 460 608**	**100 %**
average per class	7,7 %	11	343 124	7,7 %
median per class	6,7 %	7	320 594	7,2 %

Table 2: Topical categories of the texts in the corpus, with the number of books and word count in each category.

I compiled the data to be studied into a digital text corpus from online databases. As none of the currently existing corpora covered all the required texts, three different sources were used: of these, *Corpus Corporum: Repositorium operum Latinorum apud universitatem Turicensem* provided most of the texts (70 %).[53] The rest were downloaded from *PHI Classical Latin Texts* (26 %)[54] and *Musisque Deoque: A digital archive of Latin poetry* (4 %)[55]. All texts were acquired as plain text files without annotation or edited into such form.

[53] <URL: https://www.mlat.uzh.ch/home >. See also Roelli (2014). Unless otherwise indicated, texts may be downloaded for non-commercial use.
[54] <URL: https://latin.packhum.org/ >. Use is granted for personal study and copying is allowed for personal use under "Fair Use" principles of Copyright law.
[55] AA. VV., 2005, *Musisque Deoque (MQDQ)*, ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa, <URL: http://hdl.handle.net/20.500.11752/OPEN-555 >. Licensed under: Creative Commons - Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

2.1.2. Pre-processing the Texts

As will be explained below, I have analysed the texts in several formats to enable different methods. After computational and manual cleaning, the texts were lemmatized, i.e., all the inflected forms of lexemes were rendered into dictionary headwords. The most important tool in this process was *The Classical Language Toolkit* (CLTK)[56] library in Python, although a considerable amount of manual work with text editors was still required.

For quantitative approaches, I used the lemmatized texts both as plain text files and in a term-document matrix format (TDM). In the TDM, each text document possesses a row, each column represents a word (lemma), and each cell contains the frequency of the word (Aggarwal & Zhai 2012; Manning et al. 2008). The dimensions of the matrix are 255 documents x 75 063 lemmata. Semantically unimportant stopwords, such as prepositions, conjunctions, and pronouns etc. (1285 tokens in total) were removed from the matrix but retained in the lemmatized text files (Aggarwal & Zhai 2012).

To note briefly, the choice of a certain structure for a matrix is not a simple question. A major feature of TDM is that information about word order is lost when texts are presented in a bag of words form, i.e., containing only the frequencies. However, this approach is supported by successful practical applications. TDM is not the only option for representing texts as numbers but instead, it can be viewed as a special case of the more general principle of word-context matrices (Turney & Pantel 2010). In this study, TDM provides a manageable format for heterogeneous data. As both the authors and the documents represent very diverging contexts, it is illustrative to present the results at document level. In future, splitting documents into smaller units of context could enhance the resolution of analyses.

Basic information about the corpus is given in Table 3. The *quanteda*[57] package in R provided the functions for matrix formatting and, alongside other tools, for analyses. *Quanteda* also enables sampling, grouping, weighting, and filtering the data, which is useful for summarising and for performing various tests for comparison.

CORPUS, SUMMARY		PROSE	VERSE
Authors (including anonymous & uncertain)	68		
Titles	149	67 (45 %)	82 (55 %)
Text files	255	130 (51 %)	125 (49 %)
Fragmentary titles	27 (18 %)	7	20
Words (raw texts without editing), approx.	4.4 million	3.3 million	1.1 million
Word frequencies in the matrix after editing	2.8 million	2.07 million	750 000
Unique tokens (lemmata in the matrix)	75 063		
Hapax (lemmata occurring only once)	48 %	49 %	45 %
Average words per file (edited)	11 052	15 915	5 994
Median words per file (edited)	5 086	7 602	3 928
Average unique lemmata per file (edited)	2 406	3 073	1 712
Median unique lemmata per file (edited)	1 832	2 777	1 551

Table 3: Basic information about the corpus.

[56] Johnson, K. & al. (2021). The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, 20-29. <DOI: http://doi.org/10.18653/v1/2021.acl-demo.3 >.

[57] Benoit, K. & al. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* 3(30), 774. <DOI: http://doi.org/10.21105/joss.00774 >. <URL: https://quanteda.io/ >.

For qualitative analysis, I compiled a concordance (key word in context list, KWIC) of the original unlemmatized texts, centred around every occurrence of the primary target lemmata *aether, aetherius,* and *aethra* (Wynne 2008). This was to be the central way of combining qualitative and quantitative approaches (see section 2.2.2). I utilised the *AntConc* software in this, as well as in some other procedures.[58] These three lexemes occur in 112 out of 255 text files (44 %). As stated in section 2.1.1, the rest of the texts contain additional target lexemes (*astrum, sidus, stella, luna*). In the TDM, texts not containing *aether*, *aetherius*, or *aethra* simply have zero entries for these three lemmata.

2.2. Methods

2.2.1. General Notions

The objective of this study is to analyse the lexeme *aether* and its cognates semantically, and to determine their usage in the texts that have survived to our days. By examining semantic similarity or relatedness to other words, the aim is to define the semantic profile of these words as clearly as possible, i.e., which sense(s) can be attributed to them in each context (Lenci 2018). As a further goal, it is hoped that semantics could shed light on general conceptions regarding natural philosophy and world view as entertained or transmitted by ancient Roman authors.

In this study, the methodological framework chosen derives from the so-called distributional hypothesis (DH). The basic tenet of this usage-based model is that the meanings of words are defined by the contexts in which they are used: words with similar meanings tend to occur in similar contexts (Boleda 2020; Chartier & Meunier 2011; Kanner 2022; Lenci 2018). When analysing a lexeme semantically, e. g. synonymy and polysemy present significant issues (Boleda 2020). This is the principal motive for resorting to quantitative methods. By means of statistics, the researcher is in a better position to detect recurring, co-occurring, or trend-like features which could remain unnoticed by an unassisted human reader (McGillivray 2013).

Word meanings and semantic representations have successfully been extracted and validated from co-occurrence statistics (e. g. Bullinaria & Levy 2007; Günther & al. 2019). However, as Kanner (2022) recently has discussed, there are essential limitations on whether and to what extent distributions of words may be interpreted as a proxy for semantic meanings. Based on his results, the exact identity between distribution and meaning can certainly be ruled out. Results of analyses may vary significantly from one dataset to another, as the issue is apparently highly data dependent. On the other hand, there are a wide range of operations that can be implemented for semantic inquiry. Kanner (2022) stresses the importance of transparency about the chosen methods and the underlying assumptions. However, the field of methodology is far from fully trodden, as knowledge about the applicability of computational methods accumulates over time.

Although computational and statistical methods probably will not offer a shortcut to clear semantic distinctions, the stance adopted here is to use them to widen the approach and to rely on a richer set of tools. Human interpretation, while indispensable in linguistics and philology, can significantly be assisted by results gained through computational methods. If exclusive evidence cannot be attained, indicative results can act as a guide towards the right direction.

[58] Anthony, L. (2020). *AntConc*. Tokyo, Japan: Waseda University.
<URL: http://www.laurenceanthony.net/software/antconc/ >.

2.2.2. Close Reading the Concordance

In this section, I will focus on one way of combining qualitative and quantitative methods.

The concordance (KWIC, section 2.1.2) was used to devise a table for quantifying my interpretation of the lexemes *aether, aetherius,* and *aethra* in each context of usage. I read every occurrence (n = 667) of these words and interpreted it by selecting one or two out of five semantic categories. The table was inserted alongside the KWIC in such a way that a numerical value could be assigned to each occurrence. The table consists of five columns corresponding to five distinguishable semantic categories (basic meanings) derived from dictionary entries and reworded into a concise description (Table 4).

Many of the occurrences do not fit into a clear-cut semantic definition because the meanings seem to vary even within a single work, depending on the context. To take the variation and ambiguity into account, I designed the scoring scheme accordingly. An occurrence was marked either with number 1 in one of the five columns representing the semantic categories, or, if the usage was ambiguous, the value was divided between two categories (i.e., the value 0.5 was assigned to two different columns on the same row). Thus, each row adds up to one, being either 1.0 in an unambiguous case, or 0.5 + 0.5 if my interpretation was wavering between two options. Relatively many (301, 45 %) of the occurrences were labelled with a double meaning, either due to obvious ambiguity, or difficulty in interpretation.

LABEL	SEMANTIC CATEGORY	EXAMPLE TEXT
aether_AIR	Air	*si vescitur aura aetherea*[59]
aether_FIRE	Fire, heat, lightning, light	*igneus aether*[60]
aether_DISTINCT	Fifth, distinct element (not earth, water, fire, or air)	*Sed caelum ipsum stellaeque caeligenae omnisque siderea conpago aether uocatur - - elementum non unum ex quattuor quae nota sunt cunctis, sed longe aliud, numero quintum, primum ordine*[61]
aether_SKY	Heaven, sky	*grues - - tunc aethera latius implent*[62]
aether_MYTH	Mythical/abstract context (e. g. proper name, or denoting abode of the gods)	*nec tu Pittheidos Aethrae filius*[63]

Table 4: The five semantic categories (broadly, meanings) of *aether, aetherius*, and *aethra,* with exemplifying citations. The translations are mine.

The sums by columns indicate the distribution of meanings and, presumably, give information about the usage of the lexemes. The table was then inserted as five additional lemmata into the term-document matrix built from the corpus (2.1.2 and 2.2.3). Thus, it is possible to perform computational analyses with and without the researcher's interpretations: the interpreted values (i.e., the five additional columns) can be included or left out of calculations, and the results can be compared.

I have examined the distributions per time period, seeking to trace developments in usage. Unfortunately, a comprehensive view is hindered by the unequal availability of texts from different periods. In this paper, the KWIC values are mainly based on frequencies without weighting. As a

[59] "If he feeds on heavenly air". Vergilius, *Aeneis* 1.546-547.
[60] "Fiery ether". Seneca, *Naturales quaestiones* 6.16.2.
[61] "But the sky itself, the heaven-born stars, and the whole starry composition is called ether - - It is not one of the four elements known to all, but one that is very much different, fifth in order but first in rank". Apuleius, *De mundo* 1.
[62] "Cranes - - then fill the sky widely". Statius, *Thebais* 12.515-516.
[63] "Nor are you the son of Pittheus' daughter, Aethra." Ovidius, *epistulae* 10.111-112.

further step, I will weight the scores according to the proportion of each text in its respective time period's total word count. This is one way to scale the changes over time to the text's representativeness in its category.

Obviously, the role of the researcher is crucial at this phase. The interpretations of the text passages are mine, although based on established dictionary entries. This is exactly why I have implemented a quantitative notation of interpretation because it enables falsifiability in detail. In the results section, I will also exemplify one way of forming a parallel overview of the semantic categories using correspondence analysis (sections 2.2.3 and 3.2).

2.2.3. Distant Reading with Quantitative Methods

In parallel with reading and interpreting the text passages, I am using computational tools for assistance, comparison, and evaluation. These tools are implemented on the lemmatized texts, both as plain text files and as transformed into numerical matrices. The methods derive mainly from the family known collectively as the vector space model, which is based on the distributional hypothesis described in section 2.1.1 (Boleda 2020; Chartier & Meunier 2011; Günther & al. 2019; Sprugnoli et al. 2021). With methods based on DH, semantic similarity can be quantified and measured. In general, many ways of applying word (co-)occurrence statistics are available (cf. Manning & Schütze 1999). This field is markedly an intersection between corpus linguistics and data science.

I will consider two operations (cosine similarity and correspondence analysis) in this subsection. A matrix format enables the representation of documents and words as vectors, i.e., as sequences of real numbers. This, in turn, makes it possible to efficiently perform diverse calculations to extract information about the corpus (Manning & Schütze 1999; McGillivray 2013). However, absolute frequencies are usually not the best values to operate on, given the wide variance in document length. Otherwise, the outcome could be distorted in favour of longer texts. Hence, one or several modes of weighting the values usually becomes necessary. Thus, I have used the matrix in several versions that have the same dimensions, documents, and words, but in which the values differ according to the chosen weighting formula.

Using relative frequencies (percentages) instead of absolute frequencies is a slight improvement, but setting the values in proportion to the maximum or average frequencies per document will eliminate or attenuate the effect of document length. Log-average weighting for non-zero values is an informative way of representing a word's frequency compared to the average frequency in each text. (Figure 1; Manning et al. 2008). Average scaling removes the problem of text length, and using logarithms will scale down wide-ranging quantities.

$$\frac{1 + \log(tf)}{1 + \log(ave(t\epsilon d(tf)))}$$

Figure 1: Log-average weighting for word frequencies (*tf* = frequency of the word in question, *ave(tєd(tf))* = average word frequency in the document in question).

I have compared calculations with no weighting to log-average and inverse document frequency (tf-idf) weighting. Tf-idf suppresses very common words across the corpus but highlights words which occur frequently in a relatively small group of documents, making it an effective method to detect possibly relevant and semantically distinctive words (Figure 2; Chartier & Meunier 2011; Manning et al. 2008; Turney & Pantel 2010).

$$\text{idf}(t) = \log \frac{N}{\text{df}(t)}$$

Figure 2: Tf-idf weighting for documents (N = total number of documents in the corpus, $df(t)$ = the number of documents containing term t).

When seeking to define the similarity or distance between words computationally, the cosine is a common metric. It is a geometrical representation of word vectors and of the angular distance between them (Figure 3; Chartier & Meunier 2011; Manning & Schütze 1999; McGillivray 2013; Turney & Pantel 2010). This operation is generally available in NLP tools, such as *quanteda*. The cosine is a reasonable starting point, despite its averaging effect in that it, in a sense, conflates all the occurrences of a word in every context into a single value.

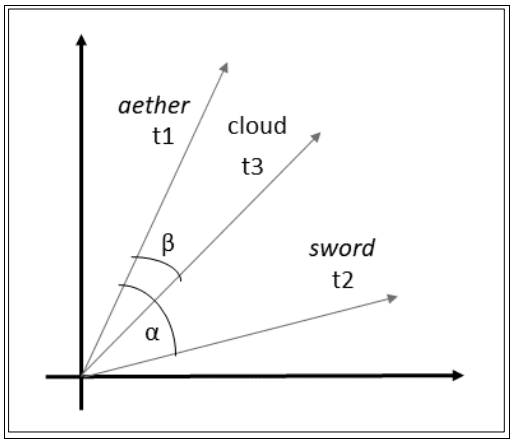


Figure 3: A schematic example of cosine similarity: α = the angle between t1 and t2 (words 1 & 2), β = the angle between t1 and t3 (words 1 & 3). In vectorial space, t1 (*aether*) and t3 (cloud) are more similar than t1 and t2 (sword).

For comparison, the interpreted values in the KWIC (section 2.2.2) were inserted as additional columns to the term-document matrix since they, too, can be operated as vectors. These interpreted values were made discernible by way of annotated lemmata. For example, the semantic category 'air' was inserted with the label *aether_AIR*, 'sky' as *aether_SKY*, etc. (Table 4). This rendered it possible to cross-check whether my interpretations and cosine calculations are along the same line, or whether there are marked differences to discern.

The semantic categories in the KWIC table were also analysed with the package *ca* in R environment, using correspondence analysis (CA).[64] CA is a quantitative method that serves as an illustrative means to summarize large amounts of data. Here, it will also help to evaluate my interpretations of the semantics of *aether*. With CA, a high-dimensional calculation can be reduced to a two-dimensional representation of associations between variables. Despite the loss of information in dimension reduction, it can serve as a descriptive tool (Agresti 2012; McGillivray 2013).

3. Results

In this section, some tentative results will be described. It should be emphasised that 85 % of the occurrences *aether* and its cognates are found in works in verse. This has an effect both on

[64] Nenadic, O. & Greenacre, M. (2007). Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package. *Journal of Statistical Software*, 20(3), 1-13. <DOI: https://doi.org/10.18637/jss.v020.i03 >.

interpretation and on some quantitative methods. Namely, Latin poetry is strictly bound to metres based on syllable length. Accordingly, word order (and choice of words) is adjusted to metre, which presents a challenge for interpretation. Computational analyses, too, are faced with great challenges because, for example, the window of observation must be wide for associated words to be properly detected, when I will later implement n-gram analyses on the texts.

3.1. Distributional Semantics of aether

The figures below are based on the KWIC table containing my quantified interpretations of the occurrences of *aether, aetherius,* and *aethra* (section 2.2.2). The percentages are based on frequencies without weightings.

The first diagram presents the proportions of each semantic category when all three lexemes are inspected together over all 112 documents (Figure 4).

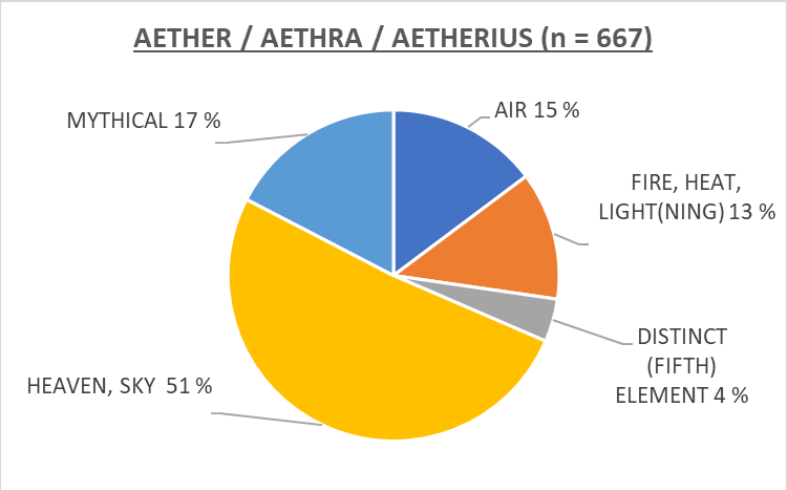


Figure 4: Percentages of the interpreted semantic categories of *aether, aetherius,* and *aethra* (no weighting).

If each lexeme is observed separately, some differences become visible, but the proportions remain mostly the same (Figures 5–7). Being the rarest word, *aethra* stands out owing to its strongly mythological contexts. It is used as a proper name 19 times, i.e., in over half of the occurrences (section 2.2.2). The adjective *aetherius* appears also to bear more mythical denotations than its root noun *aether*. For example, *pater aetherius* 'heavenly father' refers to the god Jupiter.

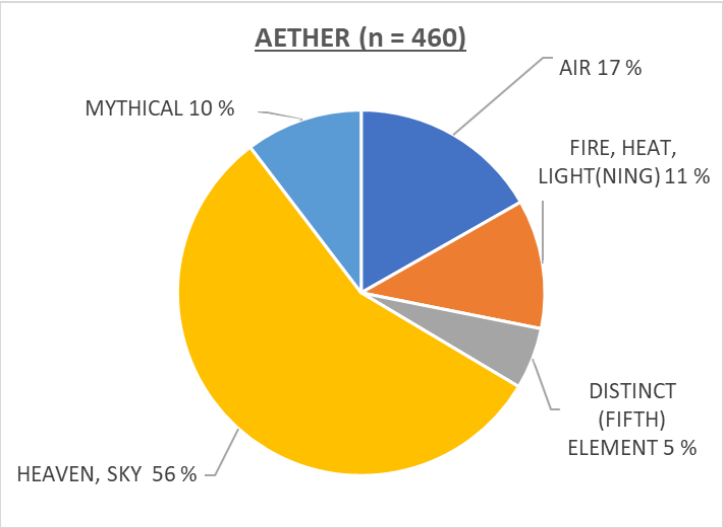


Figure 5: Percentages of the interpreted semantic categories of *aether* (no weighting).

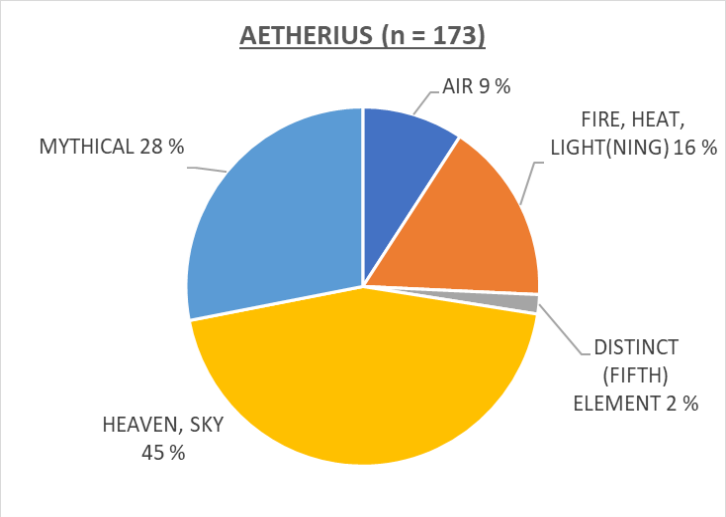


Figure 6: Percentages of the interpreted semantic categories of *aetherius* (no weighting).

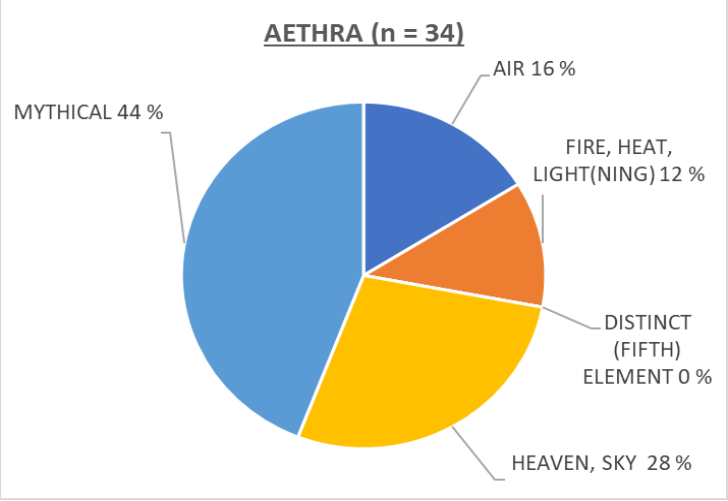


Figure 7: Percentages of the interpreted semantic categories of *aethra* (no weighting).

The distributions can be inspected diachronically if the percentages are viewed by era (Figure 8). It should be noted that, despite the linear mode of presentation chosen here, the history of the concept of *aether* cannot be straightforwardly interpreted in a similarly linear way. The lacunae in our sources leave much in the dark, so this must be taken as an indicative approximation.

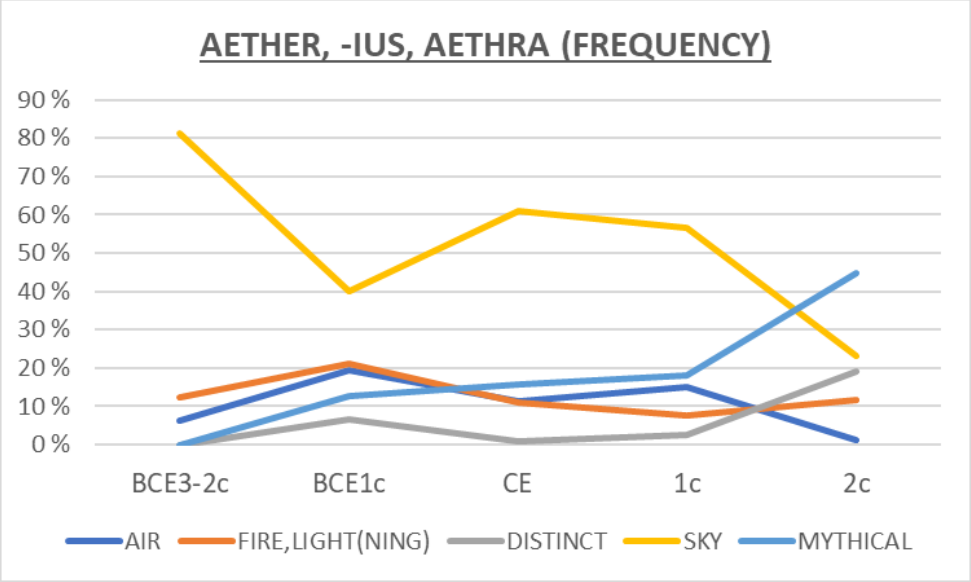


Figure 8: Percentages of the interpreted semantic categories of *aether, aetherius,* and *aethra* by era (no weighting).

As stated above, certain weighting schemes would probably enhance the informativity of this kind of presentation. This work is still in process, but a tentative version is shown in Figure 9. The calculation is performed by weighting each work by the percentage of that work in the total word count of its time period. The assumed effect is to scale the works closer to overall representativity.

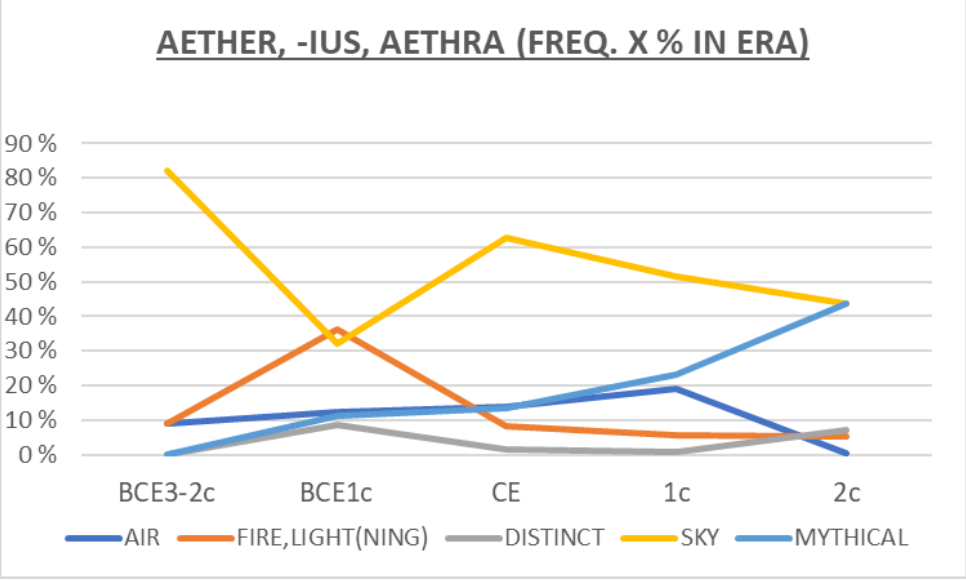


Figure 9: Percentages of the interpreted semantic categories of *aether, aetherius,* and *aethra* by era (weighted by the proportion of each work in the total word count of its time period).

This simple weighting operation does not significantly affect the apparent trends but it highlights some proportional changes. For instance, the rise in the percentage of denotations referring

to fire, heat, light, or lightning is more marked. However, the usage of *aether* and its cognates as referring to the sky is clearly the most common occurrence, perhaps matched only by the rise in mythological contexts in the second century, though there are not many works dating from that era in the corpus.

One can attempt a rough interpretation of the chronological development. First, the few archaic texts in the corpus dominantly feature *aether* as denoting the sky. Apparently, the playwright Pacuvius (ca. 220-130 BCE) even directly equated the words *aether* and *caelum* 'heaven', though this is only indirectly known through fragments. In contrast, the high percentage of 'fire' during the first century BCE might be due to Cicero's (106-43 BCE) significant role as an adapter of Greek philosophy into Latin. The overall heterogeneity probably reflects the fact that astrology steadily gained popularity in Rome, especially from the late first century BCE onwards (Soubiran 1979). Hence, both philosophical treatises and poetic depictions contain references to *aether* in various ways, along with the ever-present mythological passages and general references to the sky.

3.2. Correspondence Analysis of the Semantic Categories

The semantic categories in the KWIC table were analysed using correspondence analysis (sections 2.2.2 and 2.2.3). In the graph (Figure 10), the percentages in the horizontal and vertical axes indicate the amount of variance in the data that the results account for. The greater the percentages, the more the analysis can explain of the variance present in the data. Here, the data have been weighted by proportions of the texts in the word count across the entire corpus (not by era).

Figure 10: Correspondence analysis graph of the five semantic categories of *aether*. The blue points represent the five semantic categories, and the red points show how the individual texts are situated respective to the whole.

The resulting graph shows that dimensions 1 (51.67 %) and 2 (25.47 %) account for ca. 77 %

of the variation, which is a relatively good result because the actual calculation is high-dimensional. In general, the closer a point is to the origin, the less distinctive it is relative to the overall variance. The categories lie further apart along x-axis than along y-axis, so dimension 1 accounts for a greater part of the variance.

The meanings ‘air’ and ‘sky’ lie close together, which is in keeping with my impression while reading the passages. The connection is quite logical, as many of the double meanings in the KWIC table are indeed shared between these two senses. ‘Sky’ is also the most common, i.e., the least distinctive, denotation which explains why it is found near the origin. In CA tables, as in cosine similarity, the angles between vectors drawn from the origin to data points play an illustrative role (section 2.2.3). Small angles between vectors drawn from the origin indicate stronger association between the points, even if the lengths of vectors differ. If the vectors from the origin to the points lie roughly at a right angle, they do not show association, and opposite directions mean negative correlation.

Mythological denotations seem to dwell in their own quarter, both in the graph and in the texts, too. References interpreted as a fiery or a distinct element occur more rarely and are more distinctive, thus are further from the origin. However, at first glance it seems surprising that these two categories stand nearly at the opposite quarter compared to ‘sky’, since the contexts usually involve the heaven as a location. It might be that the second dimension, in which all five are closer (roughly within one unit), accounts partly for the fact that *aether* is practically always a reference to the outermost region of the ancient geocentric universe. A further step to examine the semantic interpretations case by case would be to focus on individual text passages (shown as the red points in Figure 10).

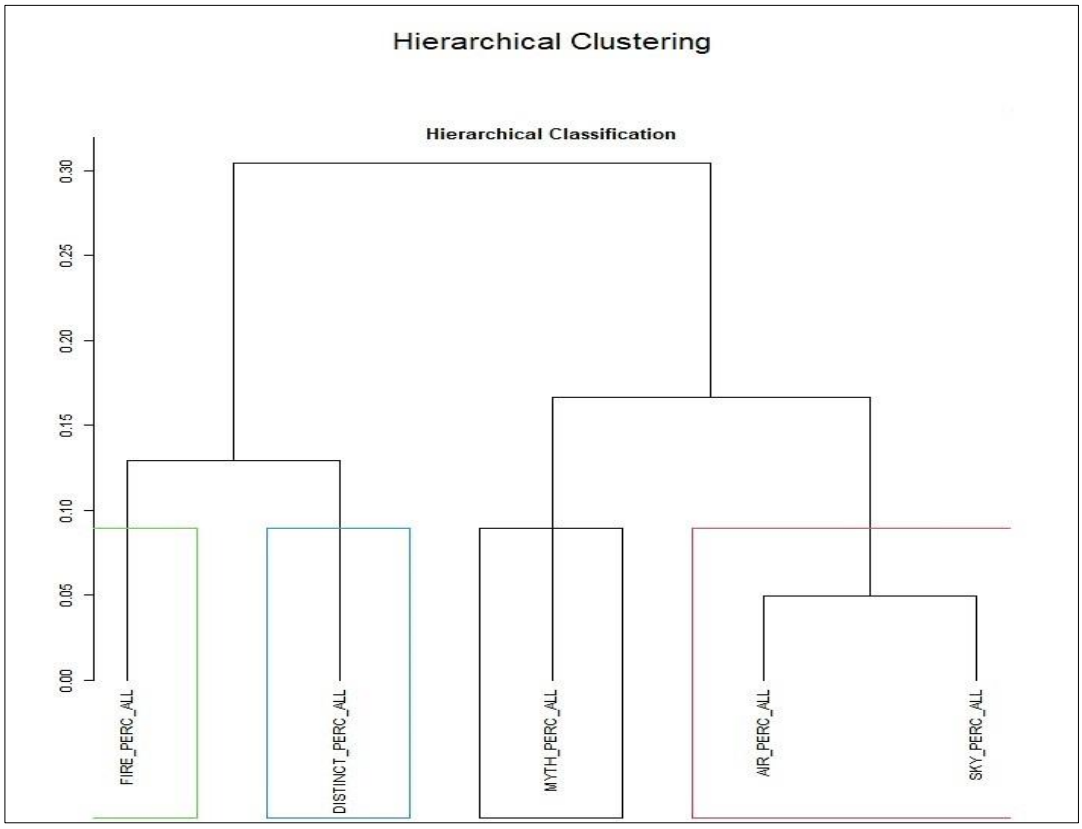


Figure 11: A hierarchical cluster graph of the five semantic categories of *aether*.

When a hierarchical cluster graph is drawn from the same data, the relations of the semantic

categories become clearer (Figure 11). On the right, the meanings ‘sky’ and ‘air’ are naturally close together, as many of the text passages could be interpreted in either way (compare, for example, the expressions ‘flying through the air’ and ‘flying in the sky’). Related to these two, the mythological or religious denotations of *aether* as the abode of the gods overlap with more general references to heaven. Finally, the material descriptions of, or allusions to, *aether* as a fiery or a completely unique element are more distant from the other three categories.

3.3. Cosine Similarity and aether

The semantic similarity of words can be roughly calculated using cosine similarity as metric (section 2.2.3). The most similar words to *aether*, as well as the effect of weighting word frequencies, can be seen in Table 5. First, it contains the most similar lexemes to *aether* with their translations and associated cosine values, calculated from the frequency matrix without weightings. The values in the middle columns are calculated from the log-average weighted matrix, where the text length is rendered irrelevant. The right-hand columns contain tf-idf weighted log-average values.

Note that the matrices also include my KWIC interpretations (labelled *aether_SKY, aether_AIR* etc.) as columns similar to other lemmata (see Table 4). Thus, it is possible to quantitatively measure the distributional similarity of *aether* to any of my interpretations, as well as to actual words occurring in the texts. As a working hypothesis, one can observe if one of the semantic categories stands out as the most similar to *aether* and, thus, could be regarded as the principal meaning. Cosine calculations can also be implemented on each of the five categories individually to detect similar lemmata. And further, by cross-checking the results and by observing the cosine values of possibly emerging groups of words, one may detect the closest synonyms and build semantic profiles of related words.

A majority of the resulting words remain the same, especially between log-average and idf columns, although with differing cosine values. Without weighting, some mythology-related words stand out. In principle, all five semantic categories are covered, perhaps apart from the meaning ‘fifth element’ which is rare. However, the cosine values should not be interpreted as absolute values semantically, but rather as indicative and relative. Since many of the scores shown in Table 5 are practically identical, relative changes between ranks matter more than the scores.

It is evident that, of the actual five categories that were annotated and added to the matrix, only *aether_SKY* and *aether_AIR* are found among the highest cosine values. This might highlight the point that ‘sky’ should be considered the general denotation of *aether*. In addition, the concepts of heaven and earth seem closely bound. When reading the texts, this becomes evident both in concrete contexts (physical descriptions) and in cases of abstract connotations (interaction between mortals and divinities). These joint occurrences could explain the fact that *tellus* ‘earth’ ranks relatively high on all three lists, as heaven and earth conceptually form a complementary pair.

FREQUENCY (no weighting)			LOG-AVERAGE			IDF x LOG-AVERAGE		
aether_AIR		0.93	*unda*	wave	0.90	***aether_SKY***		0.91
uolucer	flying	0.86	*nubes*	cloud	0.76	*unda*	wave	0.78
pharetra	quiver	0.83	*numen*	divinity	0.75	*nubes*	cloud	0.76
genitor	begetter	0.82	*flamma*	flame	0.74	*numen*	divinity	0.75
unda	wave	0.82	*tellus*	earth	0.74	*tellus*	earth	0.75
pectus	chest	0.82	*aequor*	even surface	0.74	*aequor*	even surface	0.75
artus	narrow	0.81	***aether_AIR***		0.73	*flamma*	flame	0.75
tellus	earth	0.80	*uertex*	whirl	0.71	***aether_AIR***		0.74
monstrum	portent	0.80	*aetherius*	heavenly	0.71	*aetherius*	heavenly	0.72
ingemo	to mourn	0.80	*aura*	breeze	0.70	*uertex*	whirl	0.71
nubila	cloud	0.79	*pectus*	chest	0.70	*aura*	breeze	0.71
gelidus	icy	0.79	*caelum*	sky	0.70	*pectus*	chest	0.71
altus	high	0.79	*linquo*	to leave	0.70	*caelum*	sky	0.70
crinis	hair	0.79	*fatum*	prediction	0.69	*ensis*	sword	0.70
horreo	to stand erect	0.79	*letum*	death	0.69	*sono*	to make a sound	0.70
murmur	roar	0.79	*ensis*	sword	0.68	*letum*	death	0.70
aura	breeze	0.79	*sono*	to make a sound	0.68	*linquo*	to leave	0.70
proles	offspring	0.79	*rapidus*	fierce	0.68	*fatum*	prediction	0.70
gestamen	burden	0.78	*sidus*	star	0.68	*rapidus*	fierce	0.69
aruum	field	0.78	*aruum*	field	0.68	*lumen*	light	0.69
aequor	even surface	0.78	*lumen*	light	0.68	*aruum*	field	0.69
aetherius	heavenly	0.78	*premo*	to press	0.68	*sidus*	star	0.69
sopor	deep sleep	0.78	*gelidus*	icy	0.67	*premo*	to press	0.68
sibilo	to hiss	0.78	*saeuus*	savage	0.67	*fulmen*	lightning	0.68
furio	to be mad	0.77	*altus*	high	0.67	*saeuus*	savage	0.68
aegis	shield (of Jupiter)	0.77	*fulmen*	lightning	0.67	*gelidus*	icy	0.68
umbra	shadow	0.77	*ater*	dark	0.67	*uolucer*	flying	0.68
antrum	cave	0.77	*uolucer*	flying	0.67	*altus*	high	0.68
ignis	fire	0.77	*ignis*	fire	0.66	*ater*	dark	0.68

Table 5: Most similar words to *aether* as measured by cosine from three different matrices.

4. Discussion

As for the semantics of *aether*, ambiguity seems to be typical. Double entries in the KWIC interpretation table indicate combinations of two senses, most often air and sky, or physical and abstract heaven (see section 2.2.2). When reading the actual text passages, especially the distinction between material and abstract (or religious) heaven is sometimes hard to perceive - or it may well be that the ambiguity is intentional as a literary device.

Usually, when the authors speak of ‘heavenly fires’, they obviously point to stars shining in the sky, in which case *aether* mainly acts as the locational reference (the sky). Whether it carries the idea that the region of stars is entirely composed of fire is often left ambiguous, especially in poetry, i.e., in most of the texts in which *aether* occurs. The authors may have favoured metonymical or metaphorical expressions, using air as a reference to the sky, or in general, material referring to

location: identifying the actual referent (*caelum* 'sky') by something (*aether*) associated with it.
As has been shown, 'sky' can be considered the principal denotation of *aether.* Thus, *caelum* as the primary Latin word having this meaning is naturally interchangeable with it. Noteworthily, while *caelum* crops up among the most similar words, it does not rank among the highest scores. The explanation might be that *caelum* is much more frequent than *aether* but semantically a more consistent word. Thus, their distributions are less congruent than might be expected. However, even if supposing that *aether* should be interpreted as the substance constituting the heavens, the situation could be explained with hyponymy, i.e., *caelum* as the more general word and *aether* as subordinate to it.

In contrast, clearly Aristotelian references to *aether* as a fifth, celestial element are very rare in the observed texts, if my interpretations have succeeded in grasping the original meaning. This denotation features more prominently in later sources, especially Late Medieval and Early Modern Latin texts. Therefore, the role of this sense should be put to proportions in dictionaries, historically speaking.

As for methods, the combination of manually annotated KWIC and term-document matrix holds potential. The limits, however, seem obvious: for instance, the two operations presented in this paper (correspondence analysis and cosine) are indicative rather than exact instruments - though this usually holds good for qualitative interpretation, too.

Further steps include selecting a relevant portion of the matrix to be weighted and analysed in various ways. With over 75 000 lemmata, the data is inevitably noisy. Methods and metrics I have planned to use include topic modelling, n-gram analysis, exploring collocations using pointwise mutual information, and observing the semantics of *aether* over the 13 topical categories of the texts. As the original objective of this project, the nature of the connection between *aether* and words denoting celestial bodies still needs further studies. While *aether* is a meagrely occurring lexeme, seeking similarly distributed words might shed light not only to semantics but to conceptual history of world view.

References

Aggarwal, C. & Zhai, C. (eds.) (2012). *Mining Text Data*. Springer US.

Agresti, A. (2012). *Categorical Data Analysis*. 3rd Edition. Hoboken, New Jersey: John Wiley & Sons.

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics* 6, 213-34.

Bullinaria, J. & Levy, J. (2007). Extracting Semantic Representations from Word Co-Occurrence Statistics: A Computational Study. *Behavior Research Methods*, 39 (3), 510–526.

Chartier, J. & Meunier, J. (2011). Text Mining Methods for Social Representation Analysis in Large Corpora. *Papers on Social Representations*, Vol. 20 No. 2: Special Issue: A Half Century of Social Representations: Some Recommended Papers, 37.1-37.47.

Couprie, D. (2011). *Heaven and Earth in Ancient Greek Cosmology: From Thales to Heraclides Ponticus*. New York, NY: Springer Science & Business Media.

Günther, F., Rinaldi, L., Marelli, M. (2019). Vector-Space Models of Semantic Representation from a Cognitive Perspective: A Discussion of Common Misconceptions. *Perspectives on Psychological Science* 14 (6), 1006–1033.

Kanner, A. (2022). *Meaning in Distributions: A Study on Computational Methods in Lexical Semantics*. University of Helsinki.

Le Boeuffle, A. (1987). *Astronomie, Astrologie: Lexique Latin*. Paris: Picard.

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics* 4, 151–171.

Lloyd, G. (1970). *Early Greek Science: Thales to Aristotle*. New York: Norton.

Manning, C., Raghavan, P., Schütze, H. (2008). *Introduction to Information Retrieval*. New York: Cambridge University Press.

Manning, C. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. 6th ed. with

corrections 2003. Cambridge, MA: MIT Press.
McGillivray, B. (2013). *Methods in Latin Computational Linguistics*. Leiden: BRILL.
Soubiran, J. (1979). L'astronomie à Rome. In Aujac, G. & Soubiran, J. (eds.), *L'Astronomie dans I 'antiquité classique. Actes du Colloque tenu à l'Université de Toulouse-de-Mirail, 21-23 octobre 1977*. Paris, 167–183.
Sprugnoli, R., Moretti, G., Passarotti, M. (2021). Building and Comparing Lemma Embeddings for Latin. Classical Latin versus Thomas Aquinas. *Italian Journal of Computational Linguistics*, 6(1), 29-45.
TLL = *Thesaurus Linguae Latinae* (1900-). Lipsiae: B. G. Teubner.
Turney, P. & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37, 141–188.
Wright, M. (1995). *Cosmology in Antiquity*. London: Routledge.
Wynne, M. (2008). Searching and Concordancing. In Kytö, M. & Lüdeling, A. (2008). *Corpus linguistics: An International Handbook*. Volume 1. Berlin: Mouton de Gruyter.
Zucker, A. (ed.) (2016). *L'encyclopédie du Ciel: Mythologie, Astronomie, Astrologie*. Paris: Robert Laffont.

Building the Corpus of Finland-Swedish Sign Language: Acknowledging the Language History and Future Revitalization

Juhana Salonen[1], Maria Andersson-Koski[2], Karin Hoyer[3], Tommi Jantunen[4]

University of Jyväskylä1, 3-4, University of Helsinki2

E-mail: juhana.salonen@jyu.fi

Abstract

This paper presents the first steps in the process of creating a multimedia corpus for the severely endangered Finland-Swedish Sign Language (FinSSL). In the paper, we will first outline the history and current situation of FinSSL and then move on to describe some of the foundational choices which we have made both in the earlier data collection and at the start of the currently ongoing annotation work. Finally, we will bring up challenges related to the corpus data processing and discuss the future uses of the corpus, especially from the point of view of the FinSSL revitalization process.

Keywords: Finland-Swedish Sign Language, corpus, annotation, Signbank, research, revitalization

1. Introduction

This paper presents the first steps in creating a multimedia corpus for Finland-Swedish Sign Language (FinSSL) and outlines how corpus-building efforts can support deaf community aims for language conservation and revitalization. Building the corpus consists of data collection, processing and annotating the data, and developing a lexical database, Signbank[65] (see Takkinen et al. 2020). There are two national sign languages in Finland: FinSSL and Finnish Sign Language (FinSL). FinSSL is a severely endangered sign language. It is used by approximately 90 deaf people in the coastal areas of Finland. The Finnish Government is currently committed to the revitalization of FinSSL, whereby the Ministry of Culture and Education has assigned the University of Jyväskylä and the University of Helsinki a shared responsibility for 2021–2024 to carry out research on FinSSL. At the University of Helsinki, the research focuses on FinSSL users' role in the process of language revitalization. At the University of Jyväskylä, the mandate is being fulfilled by building the Corpus of Finland-Swedish Sign Language (Corpus FinSSL). The Corpus FinSSL will accompany the larger Corpus of Finnish Sign Language (Corpus FinSL; see Salonen, Kronqvist & Jantunen 2020), which was partly published in the FIN-CLARIN consortium's Language Bank in 2019 [66]. Corpus FinSSL will be stored mainly at the University of Jyväskylä and will later be transferred to the FIN-CLARIN's Language Bank for long-term preservation and publication according to language informants' research consents and data protection regulations.

2. Background on Finland-Swedish Sign Language

The history of FinSSL reaches back to 1846, when Carl Oscar Malm founded the first school for the deaf in Porvoo (in Swedish, *Borgå*), Finland. At the school, Malm taught his deaf pupils with sign language he had learned in Stockholm, Sweden, and as the number of students and schools grew, so the use of Malm's sign language also spread in Finland. Due to the oralistic trend, focusing on the development of speech articulation in deaf education, early on, deaf pupils were separated into Swedish and Finnish deaf schools according to their family background, which caused Malm's sign language to diverge into two different varieties, FinSSL and FinSL (Salmi & Laakso 2005).

The existence of two signed varieties was made visible for the first time in linguistics by

[65] The University of Jyväskylä, Sign Language Centre (2019)
[66] The University of Jyväskylä, Sign Language Centre (2018)

Rissanen (1985), who described FinSSL as one of two "main dialects" in FinSL. The signing of the Finland-Swedish deaf, as she puts it, *clearly differs from the signing of the Finnish deaf* (Rissanen 1985, 14). At the time Rissanen made her observation, the school in Porvoo was the only remaining school for pupils from Finland-Swedish homes and thereby an important linguistic environment for what would later on be defined as a Finland-Swedish deaf community (Lindberg 2021a). At the end of the 1980s there was a growing discontent with how education at the Porvoo school was arranged which resulted in a decreasing number of students and finally in the closing of the school in 1993. This political decision made by the government to stop providing education for FinSSL users again contributed heavily to the migration of language users to Sweden (Lindberg 2020; 2021b). Finland-Swedish deaf who stayed in Finland either got integrated among hearing Swedish-speaking pupils or attended Finnish deaf schools (Londen 2004).

A growing awareness of the loss of a linguistically and culturally important environment due to the closing of the Porvoo school contributed to supporting measures within the community. As early as in 1981, a Swedish[67] working group was established within the Finnish Association of the Deaf (FAD) (Wallvik 2005) and in 1998–2002 the first language documentation and description project of FinSSL was carried out within FAD. The project resulted in the publication *Se vårt språk! Näe kielemme!* [Eng. See Our Language] (Hoyer & Kronlund-Saarikoski 2002), that demonstrated characteristic features of the lexicon of FinSSL that differed from FinSL. At the same time, in 2002, a separate club for Finland-Swedish signers, called *Finlandssvenska teckenspråkiga rf (FST)*, was founded. In 2005, FST gave a response to the inquiry by the Ministry of Justice for the *Government Report on Application of Language Legislation* and declared their language to be a language of its own (Hedrén et al. 2005).

The definition of FinSSL as a separate language was also a precondition for introducing the context of language revitalization. Since the beginning of the 21st century, FST, together with FAD, have played an important role in providing information and promoting linguistic rights for FinSSL users. A more explicitly formulated step toward language revitalization was preceded by answering the adapted UNESCO survey on endangered sign languages (Safar & Webster 2014). FinSSL was labeled a *severely endangered language* in 2013. The scoring of vitality according to UNESCO criteria emphasized the gravity of the language situation and had an impact on decision-makers. In 2015, FinSSL was recognized in the Sign Language Act (Viittomakielilaki 359/2015), and at the same time the Finnish Government granted project funding for language revitalization.

Today, FinSSL is used by approximately 90 deaf people in the coastal areas of Finland. The total number of FinSSL users is, however, estimated to be somewhat higher, since the language is also used by hearing people (Andersson-Koski 2015). The shared research responsibility of the universities is a part of the ongoing government-funded revitalization.

3. Corpus Work on Finland-Swedish Sign Language

3.1. Collecting the Data

The video data for the Corpus FinSSL was collected alongside the Corpus FinSL data from 2015–2017 with the help of one Finland-Swedish deaf signer managing the recording sessions with informants. The material was recorded in both Jyväskylä (the University of Jyväskylä's television

[67] The term *Finland-Swedish sign language* was not yet established in the 1980s. Instead, the group was referred to as the "Swedish deaf" or "Swedish-speaking deaf" whose language was characterized by signs identified as typical for the school in Porvoo (swe: *Borgåtecken*) (Hoyer 2005).

studio) and Helsinki (FAD's studio). The FinSSL data contains elicited narratives and conversations from 12 FinSSL signers aged between 28 and 89 years, of whom there were 7 men and 5 women and most of whom live in Southern Finland. Seven task types were used in the data collection (see Table 1).

	Task type	**Description**
1.	Presenting oneself	Signers present themselves and tell briefly about their background.
2.	Telling about one's hobby/work	Signers discuss their work history or hobby.
3.	Signing cartoon strips	Retelling the contents of 4 frames of *Ferd'nand* cartoon strips.
4.	Signing a video story	Retelling the contents of short *Mr. Bean* and *Laurel & Hardy* movies.
5.	Signing from a picture book	Retelling the contents of textless picture books *The Snowman* and *Frog, Where are you?*
6.	Discussing an event related to Deaf culture	Signers have a conversation about an event that is related to Deaf culture and which they have personal experience of.
7.	Free discussion	Signers discuss a topic of their choice.

Table 1: The task types of data collection.

The data was recorded using six to seven high-quality Panasonic video cameras (3 x AG-HPX371E, 1 x AW-HE120KE, 3 x AG-HPX171E). The total duration of the video data is approximately 50 hours (including all camera angles). The signers participated in the tasks in pairs. Camera 1 recorded a general view of the both signers and cameras 2 and 3 recorded full frontal views of both signers (Signer A/B). Cameras 4 and 5 were focused on the torso and face of signer A and B, respectively. Camera 6 was located in the ceiling directly above both signers for getting exact information on the different body parts on the sagittal plane (this camera angle was not used in Helsinki). The last camera 7 recorded the instructor (see Figure 1). The HD recordings were saved on P2-disks (25–50 fps), stored in MXF format and compressed into MP4 files. (see Salonen et al. 2016.)

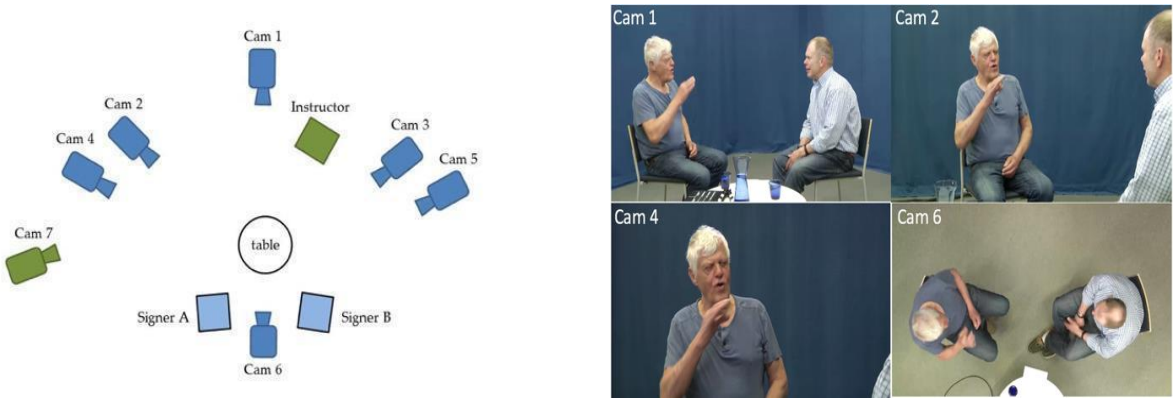


Figure 1: Camera setting in the recording (a) (Salonen et al. 2016); data from different angles (b).

During the recording sessions, we also collected consent information and metadata from the signers. On the consent forms, the signers were asked for permission to use their signing in the video material for research, display and publication purposes. All signers also had the option to not give their consent. On the metadata forms, the signers were asked about their personal, family and language backgrounds (e.g., region of residence, age). The final metadata information also includes technical details about the material and its collection (e.g., the size of the corpus, materials used in the elicitation tasks, etc.). A more detailed description of the consent and metadata forms is available in Salonen, Kronqvist & Jantunen (2020).

3.2. Annotating the Data

The annotation work of the FinSSL videos began in January 2021. The basic annotation was designed to follow the model used in the annotation of Corpus FinSL. This means the video data is processed with sign-level ID-glosses and sentence-level translations using ELAN annotation software (ELAN 2022; Crasborn & Sloetjes 2008), with Swedish as the written metalanguage.

3.2.1. Sign-Level Annotation

Building a functioning corpus demands unity and consistency with common principles and annotation guidelines for sign tokens (see Keränen et al. 2016). Many sign language corpus projects have developed their own annotation conventions (e.g., Schembri et al. 2013 for the United Kingdom; Crasborn et al. 2015 for the Netherlands; Johnston 2016 for Australia; Wallin & Mesch 2021 for Sweden). In Finland, the annotation process involves first the identification of sign units from the video (Jantunen 2015) and then tagging these units with ID-glosses (Salonen, Kronqvist & Jantunen 2020), which are unique form-meaning pairings that roughly represent the lemmas of traditional dictionaries. The ID-glosses are organized within the lexical online database, Finnish Signbank which is connected to ELAN over the Internet.

According to Johnston (2008; 2010), ID-glosses are tags that refer to sign tokens (both homonymous and polysemous) that all have the same form. Some phonetic variation is allowed (e.g., allophonic variation in some of the main parameters of the sign) and this is described in the Signbank entries. In practice, the ID-gloss functions as an identifier agreed upon by the annotators and enable systematic searches of the corpus data to be carried out. For example, the ID-gloss WAIT refers to homonymous sign tokens that carry the meaning ‘wait’ and ‘satisfied’ in FinSSL. In addition, ID-glosses contain provisional information about the grammatical features of the sign (e.g., repetition) if necessary. Examples of our current annotation conventions are given in Tables 2 and 3. Variants of signs are distinguished with codes written in parentheses after the main gloss of the signs (capital letters refer to different handshapes). Grammatical type or behavior is indicated with the symbol @ [where, e.g., @upprep refers to the repetition of the movement during the sign].

Category	Example
Lexical signs	a common/distinct ID-gloss:
Phonetic variants (1–2 different parameters)	LUCK(2g), LUCK(T) **=> LUCK**
Lexical variants (2–4 different parameters)	with e.g., a handshape code: **MEAN(GB) vs. MEAN(VV)**
Polysemic signs	WANT, HOPE, THINK, GOOD-MOOD => **WANT**
Homonym signs	WAIT, SATISFIED => **WAIT**

Table 2: Examples of the glossing conventions for different types of signs.

Type of grammaticality	Code
Negation	@neg
Repetition+plural	@upprep
Compound sign	@ssg
List buoy	@bojl
Lexicalized fingerspellings	@bt
Language contact	@sk

Table 3: Examples of the symbols for grammatical and usage-based features.

Finnish Signbank[68] is the lexical database built for FinSL and FinSSL corpus work. Signbank includes ID-glosses that are used for annotation in ELAN[69] as well as the citation form of the sign on video(s), the sign's Swedish equivalents, and any further information on the sign (Figure 2). The purpose of the database is to support annotation work with the help of the external controlled vocabulary (ECV). During annotation, it is possible similarly to see both ID-glosses (a left column) and their translation equivalents (a right column) in ELAN (Figure 3). This controlled feature allows the annotation work to proceed systematically, avoiding, among other things, spelling mistakes. Moreover, manual changes of ID-glosses and their equivalents in Signbank are automatically updated in ELAN, which also helps to control wide corpus data efficiently.

[68] The University of Jyväskylä, Sign Language Centre (2019)
[69] ELAN 2022

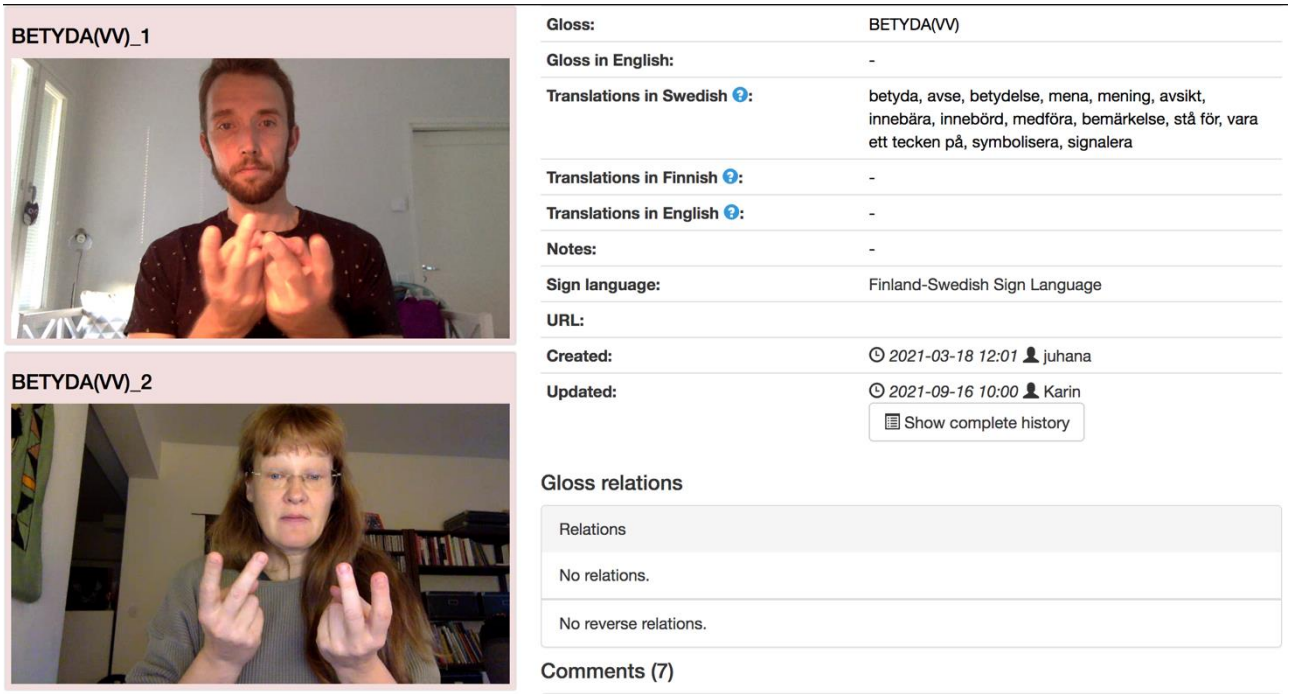


Figure 2: View of an ID-gloss working version page in Signbank. The videos on the left show the citation form(s) of the sign(s). The sign´s translational equivalents in Swedish and other information (e.g., the log of changes; the sign´s relation to other signs etc.) are described with text on the right.

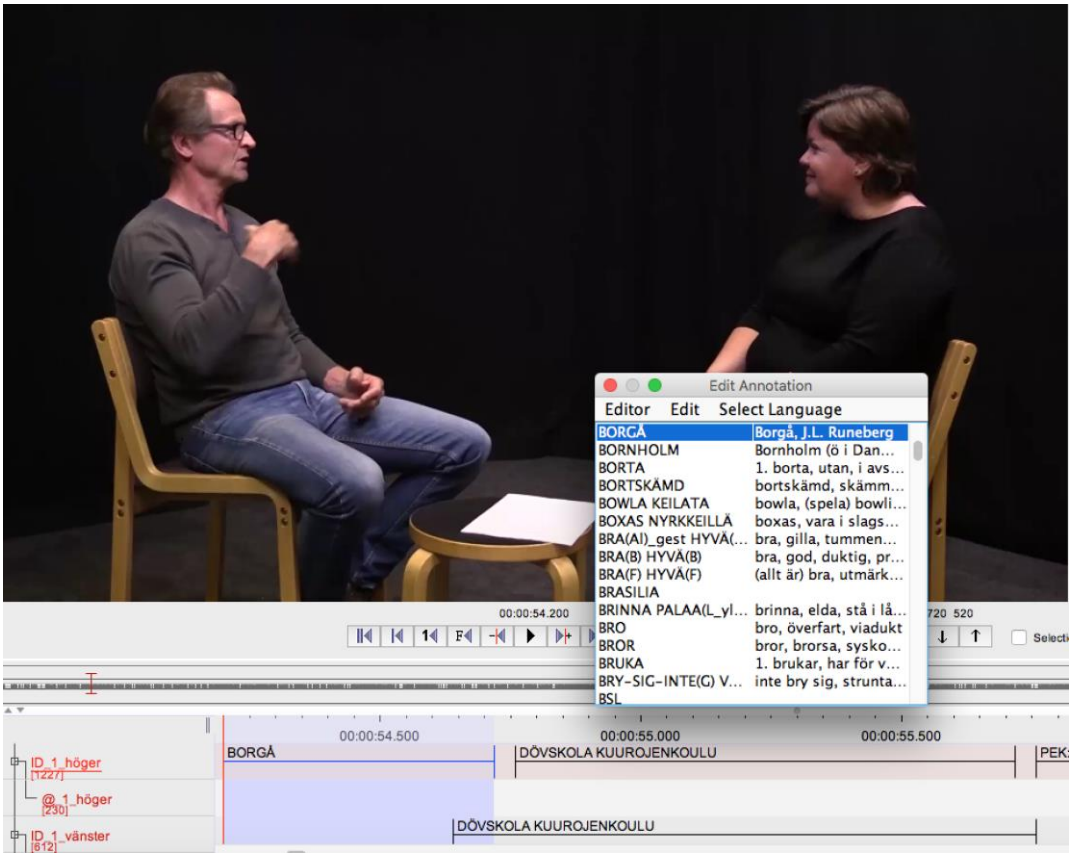


Figure 3: The view includes annotation tiers (with red font) on the left, and ID-glosses in Swedish and Finnish. The ID-glosses can be chosen from the Signbank database through the window that opens up in the view. The first column presents the ID-glosses in alphabetical order and the second column their translation equivalents in Swedish.

3.2.2. Sentence-Level Annotation

The FinSSL signing has been translated into Swedish at the level of sentences. The translation covers the meaningful information conveyed both by the manual (hands) and non-manual (other body parts) articulators. In addition, the translation seeks to distinguish sign language from the written Swedish language, so mandatory expressions in Swedish (e.g., the subject of the sentence, a copula, some conjunctions, adpositions; see Example 1) have been added in parentheses. The translation guidelines will be described in more detail in the forthcoming annotation conventions (cf. the convention for FinSL in Salonen et al. 2019).

(1) EGEN:min NAMN l-e-n-a_bokst

Mitt namn (är) Lena. [My name (is) Lena.]

The translation helps the corpus user to get a more complete view of the signed text because ID-glosses focus only on manual articulation. With the help of translations, users are able to see more accurately what meaning the ID-gloss refers to in the context. Similarly, the translation process provides support for creating Swedish translation equivalents of ID-glosses in Signbank. (cf. Salonen, Kronqvist & Jantunen 2020.)

3.3. Developing Annotation Guidelines for FinSSL

While the design of the Corpus FinSSL has benefited from earlier work on Corpus FinSL, we have also tailored our FinSSL corpus processing practices to better suit the needs of the FinSSL data and research agenda. An example is the annotation system which indicates lexical relations between FinSSL and FinSL in Signbank: we added codings for language contact, which differs from the way Corpus FinSL signs have been annotated. For example, in the FinSSL corpus data we have found three different signs that can carry the meaning ‘personal’. These signs have been given the ID-glosses PERSON(BB), PERSON(Lc), and PERSONLIG. The ID-gloss PERSONLIG is coded with @sk (språkkontakt, i.e., language contact), since the sign is a common sign in FinSL (HENKILÖKOHTAINEN) and its´ form reflects influence from Finnish (Figure 4).

At the beginning stage of annotating FinSSL, Corpus FinSL data was exploited by applying its ID-glosses and annotation guidelines. In practice, this meant that the annotator (a native FinSL signer) labeled FinSSL signs which had a similar form as in FinSL with the same ID-gloss which already existed in a FinSL lexicon of Signbank. It should be noted that the meaning may vary despite the similar form of the hands. The annotator marked separately on its own tier those sign utterances whose form varied from FinSL. Another worker, who has Swedish as her mother tongue and herself belongs to the cultural minority of Finland-Swedes, was responsible for the work on defining the form of the ID-glosses in Signbank, and on specifying the translation equivalents in Swedish (for the sign in question). At this stage, cooperation with FinSSL language guidance of various parties was emphasized (more on this in the next section).

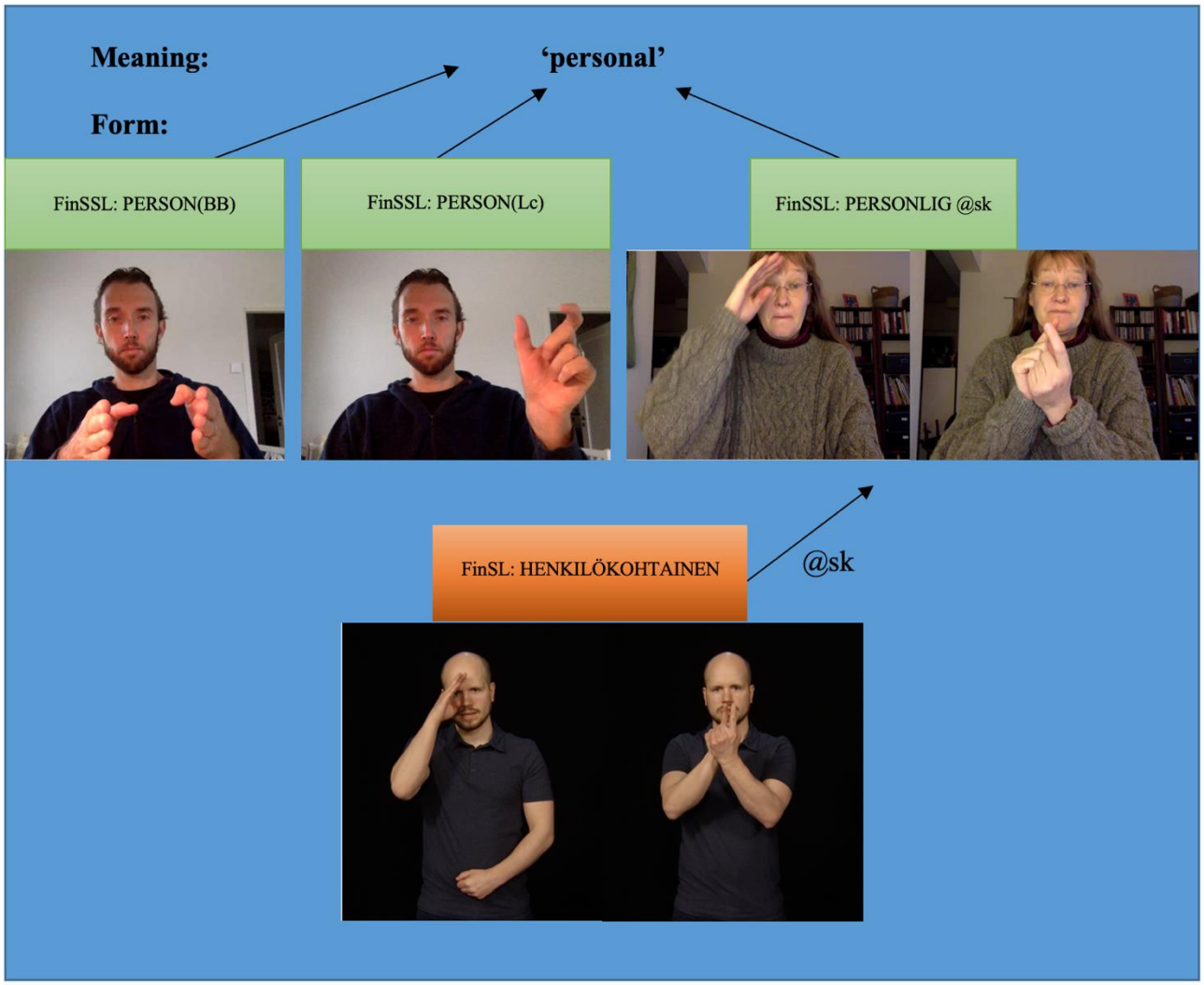


Figure 4: Example of language contact in FinSSL.

The annotation guidelines will be finalized during building the corpus, and updated over time as the scope of the annotation expands in the future as the growing research on FinSSL progresses. Similarly, updating a FinSSL lexicon in Signbank with new signs, translation equivalents and other information will serve the purpose of the FinSSL corpus. The aim is to create a completely independent entity for the FinSSL corpus, in which case the intervention of the Finnish language as well as FinSL will be eliminated. Their presence at this moment is due to solely technical regulations. In the final version, the Swedish language is used in ID-glosses (Figure 5), translation equivalents and Signbank's user interface and metalanguage.

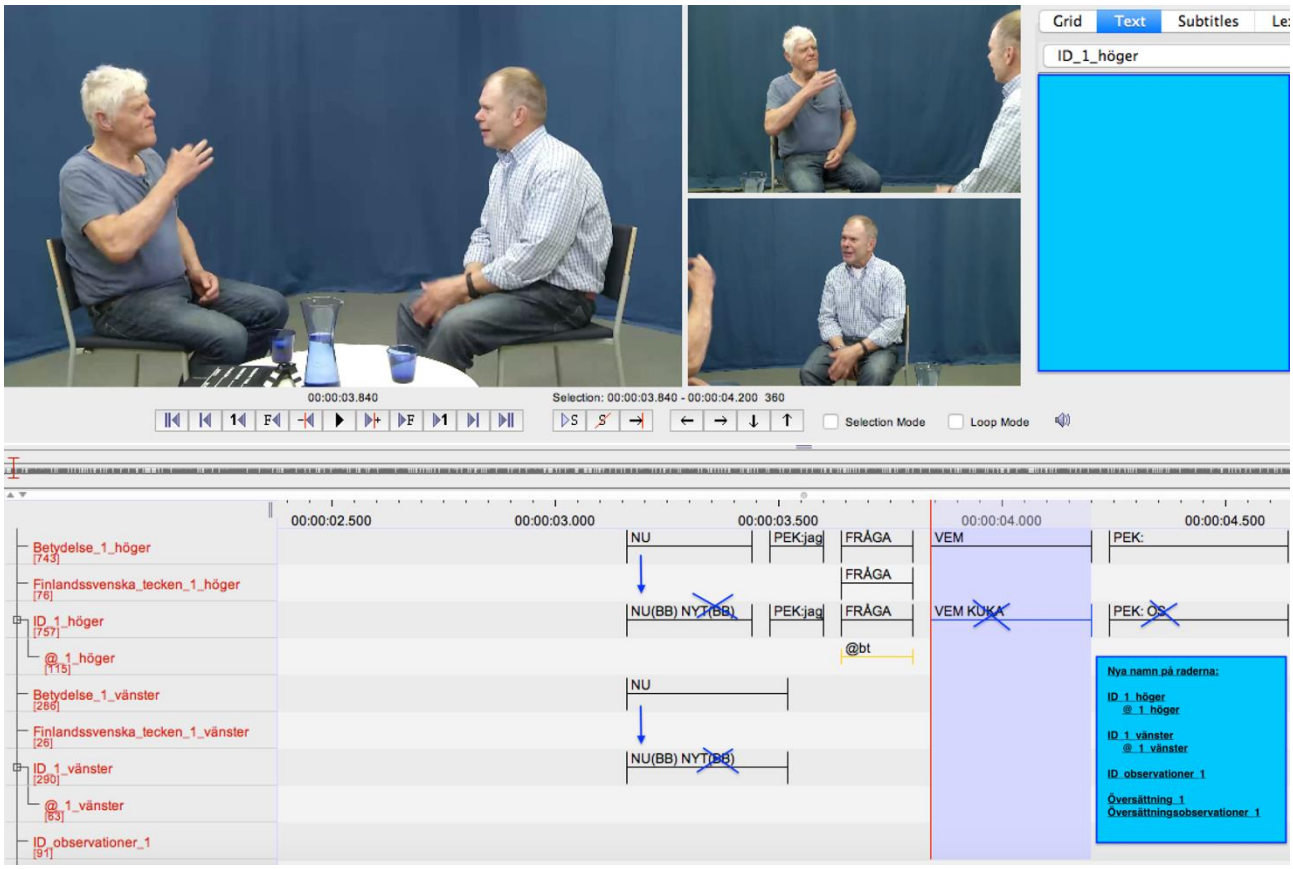


Figure 5: The view includes annotation tiers on the left, and ID-glosses in Swedish and Finnish. The process of eliminating Finnish as a metalanguage is shown by the crossed out glosses.

4. Involving the FinSSL Community – Corpus as a Tool for Empowerment and Revitalization

Certain challenges often surround work with an endangered language such as FinSSL (see Andersson-Koski 2022) which also affects the starting position for corpus work. A major challenge is the lack of previous research knowledge on FinSSL. This challenge arises in the corpus work in the need for categorization (see 3.2.1 Sign-Level Annotation), despite the lack of research on the grammar of the language. In addition, the threat of language attrition due to a loss of linguistic domains appears in the corpus data as a varying extent of language contact influences from both FinSL and Swedish Sign Language (SSL). The vulnerability of the language brings up to date the question of how to define FinSSL today (Hoyer 2012; 2013a). A further challenge in the corpus work is the lack of human resources with both linguistic and cultural knowledge in FinSSL, as none in the corpus staff represent a native FinSSL-user[70].

All these challenges contribute to emphasizing the importance of involving language users in the work process and spell out the relation between the research staff and the participants representing the language community. This calls for a need to scrutinize one's framework, concepts and starting points in the field of Deaf Studies (see Kusters, De Meulder & O'Brien 2017). To achieve linguistic guidance from native language users, the FinSSL corpus staff have included regular consultation with the FinSSL linguistic advisors of FAD. Moreover, workshops with language users have been

[70] This is due to the fact that native FinSSL-users are mainly above working age (Rainò & Vik 2020: 84), and poor availability of educational options has led to lack of academically educated language users who could be potential research staff.

implemented and will be arranged on a regular basis during the process of building the corpus. The involvement of the community, however, not only reaches toward the improved quality of the corpus itself, but highlights the relevance of its broader revitalization perspective. It is one way to engage the community to discuss FinSSL on a meta level which, eventually, will contribute to enriching the language and its use on both the individual and societal levels, both crucial to language revitalization.

The published corpus will enable and support research, teaching, lexicographic work and other activities crucial for strengthening the language. In addition, the process of creating the corpus has a significant symbolic value that improves the status enhancement of FinSSL. The fact of FinSSL being the target of academic research gives recognition of the existence of the language itself. But even more, witnessing academic appreciation for FinSSL is empowering and contributes to strengthening the linguistic and cultural identity of the community. The further acknowledgment of the language might lead to an "attitude shift" (Sallabank 2013, 65) toward the language among the surrounding majority (i.e., FinSL, Swedish and Finnish) language users. Due to the lack of research on the structure of the language, the role of the corpus as a documentation of how FinSSL looks today is emphasized. It can serve as a "mirror" for language users, who have not received any education or training in the subject of their own first language. Familiarizing oneself with the corpus even during the process of its creation, raises linguistic awareness and can be a tool for reflection and introspection about someone's own language use (see Hoyer 2013b).

In addition to its symbolic value, the engagement of language users in the process adds an educative dimension whereby users are provided with information on linguistic matters and are inspired to metatalk about their linguistic resources. Information given as a part of workshops thereby supplies more concrete tools (i.e., terms) to talk about linguistic features instead of operating with the elusive concept of "linguistic intuition." Simultaneously, it is important to bear in mind the restricted characteristics of the corpus. It represents only a snapshot of a language in transformation at a specific time used by a limited number of signers. Therefore, it cannot be set as the normative base for producing educational material without further research. However, a greater understanding of the meaning and possible use of a corpus may encourage in-depth cooperation. Members of the language community might become inspired to attend further linguistic training. This extended experience of ownership of the corpus raises its value for the community.

5. Conclusion

In this paper, we have described the process of building the multimedia corpus of FinSSL. This comprises collecting the data, consents and metadata as well as annotating the data and developing conventions for the annotation and creating a corpus lexicon. In addition to the actual corpus work, we have also presented the history and current situation of FinSSL, and the role of corpus work in the process of language revitalization.

Corpus FinSSL will be stored mainly at the University of Jyväskylä and will also be transferred to the FIN-CLARIN's Language Bank for long-term preservation and publication. The corpus of FinSSL will make it possible to promote research on the linguistic and cultural aspects of FinSSL in a comprehensive way and more systematically, even though the data in the corpus is not very wide. Similarly, the electronic and computer-readable material offers new opportunities for research on different sign languages comparatively. This is about to begin, for example, at the Nordic level in the Nordic Signed Language Corpus Network (NSLCN)[71] between FinSSL, FinSL, Norwegian Sign

[71] https://www.jyu.fi/hytk/fi/laitokset/kivi/opiskelu/tutkinto-ohjelmat-ja-oppiaineet/viittomakieli/nslcn

Language and SSL. Moreover, the FinSSL corpus material is already being used by different parties for their studies and research.

FinSSL is considered to be a severely endangered language. In addition to describing the ongoing corpus-building work, we have raised ethical issues in terms of research positionality during this work. The fact that the corpus staff are not native FinSSL signers opens up the necessity, but also the opportunity, to actively engage the FinSSL community into the working process. This has been achieved in the form of regular consultation with the FinSSL linguistic advisors of FAD. Moreover, workshops with FinSSL users and cooperation with other Finland-Swedish stakeholders have been implemented and will be arranged at regular intervals. The corpus will have a significant impact on the FinSSL community and the social status of sign language. In addition, the process of building the corpus together with the language community is already creating opportunities for strengthening linguistic awareness and confidence. This is crucial for the survival and revitalization of FinSSL as an endangered sign language.

References

Andersson-Koski, M. (2015). *Mitt eget språk – vår kultur. En kartläggning av situationen för det finlandssvenska teckenspråket och döva finlandssvenska teckenspråkiga i Finland 2014–2015*. Helsingfors: Finlandssvenska teckenspråkiga r.f.

Andersson-Koski, M. (2022). Utmaningar och lösningar för att revitalisera ett språk med liten språkmiljö – exemplet finlandssvenskt teckenspråk. In K. Kvarfordt Niia (ed.), *Framgång för små språk En översikt om varför små språk i Norden behöver stärkas och vad som bidrar till ett lyckat språkstärkande arbete*. Uppsala: Institutet för språk och folkminnen, 58–63.

Crasborn, O., Bank, R., Zwitserlood, I., Kooij, E., Meijer, A. & Sáfár, A. (2015). *Annotation Conventions for The Corpus NGT. Version 3*. Radboud University Nijmegen: Centre for Language Studies & Department of Linguistics.

Crasborn, O. & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In O. Crasborn, E. Efthimiou, T. Hanke, E. D. Thoutenhoofd & I. Zwitserlood (eds.), *Proceedings of the Sixth International Language Representation and Evaluation Conference (3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Signed Language Corpora).* May 26th –June 1st 2008. Marrakech: European Language Resources Association (ELRA), 39–43.

ELAN [Computer software] (2022). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. https://archive.mpi.nl/tla/elan

Hedrén, J., Hoyer, K., Londen, M., Wenman, L., Westerholm, H. & Östman, J.-O. (2005). Finlandssvenska teckenspråkiga i dag. In J.-O. Östman (ed.), *FinSSL – Finlandssvenskt teckenspråk*. Nordica Helsingiensia 4. Helsingfors: Nordica, 113–122.

Hoyer, K. (2005). "Vi kallade dem Borgåtecken". Det finlandssvenska teckenspråket i går och i dag. In J.-O. Östman (ed.), *FinSSL – Finlandssvenskt teckenspråk*. Nordica Helsingiensia 4. Helsingfors: Nordica, 21–80.

Hoyer, K. (2012). *Dokumentation och beskrivning som språkplanering – perspektiv från arbete med tre tecknade minoritetsspråk.* Nordica Helsingiensia 29. Helsingfors: Nordica.

Hoyer, K. (2013a). Suomenruotsalainen viittomakieli – attritiosta ahdinkoon. In K. Granqvist & P. Rainò (eds.), *Rapautuva kieli. Kirjoituksia vähemmistökielten kulumisesta ja kadosta*. Vantaa: Suomalaisen Kirjallisuuden Seura, 232–251.

Hoyer, K. (2013b). *Language vitalization through language documentation and description in the Kosovar Sign Language Community*. Nijmegen: Ishara Press. www.oapen.org/search?identifier=442947

Hoyer, K. & Kronlund-Saarikoski, K. (eds.) (2002). *Se vårt språk! – Näe kielemme! Finlandssvenskt teckenspråk 38 ordboksartiklar*. Helsingfors: Finlands Dövas Förbund r.f. & Forskningscentralen för de inhemska språken.

Jantunen, Tommi (2015). How long is the sign? *Linguistics* 53 (1), 93–124.

Johnston, T. (2008). Corpus linguistics and signed languages: No lemmata, no corpus. In O. Crasborn, E. Efthimiou, T. Hanke, E. D. Thoutenhoofd & I. Zwitserlood (eds.), *Proceedings of the Sixth International Language Representation and Evaluation Conference (3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Signed Language Corpora)*. May 26th –June 1st 2008. Marrakech: European Language Resources Association (ELRA), 82–87.

Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics* 15 (1), 106–131.
Johnston, T. (2016). *Auslan Corpus Annotation Guidelines*. Centre for Language Sciences, Department of Linguistics, Macquarie University (Sydney) and La Trobe University (Melbourne).
Keränen, J., Syrjälä, H., Salonen, J., & Takkinen, R. (2016). The Usability of the Annotation. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, & J. Mesch (eds.), *Workshop Proceedings: 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining / Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Paris: European Language Resources Association (ELRA), 111–116.
Kusters, A., De Meulder, M. & O'Brien, D. (2017). *Innovations in Deaf Studies: The role of Deaf scholars*. Oxford: Oxford University Press.
Lindberg, H. (2020). Att värna om en minoritet inom en minoritet: Finlandssvenska dövas gränsposition och arbete för rättigheter i 1980-talets Finland. *Historiska Och Litteraturhistoriska Studier* 95, 191–217. https://doi.org/10.30667/hls.87662
Lindberg, H. (2021a). National Belonging Through Signed and Spoken Languages: The Case of Finland-Swedish Deaf People in the Late Nineteenth and Early Twentieth Centuries. In V. Kivimäki, S. Suodenjoki, & T. Vahtikari (eds.), *Lived Nation as the History of Experiences and Emotions in Finland, 1800–2000* (pp. 217-239). (Palgrave Studies in the History of Experience). Cham: Palgrave Macmillan, 217–239. https://doi.org/10.1007/978-3-030-69882-9_9
Lindberg, H. (2021b). Kohti sivistystä – suomenruotsalaisten kuurojen maastamuutto Ruotsiin 1900-luvun loppupuolella. In M. Tervonen & J. Leinonen (eds.), *Vähemmistöt muuttajina: Näkökulmia suomalaisen muuttoliikehistorian moninaisuuteen*. Turku: Siirtolaisuusinstituutti. https://urn.fi/URN:ISBN:978-952-7399-10-1
Londen, M. (2004). *Communicational and educational choices for minorities within minorities: The case of the Finland-Swedish deaf*. Research Report 193. Department of Education. Helsinki: Helsinki University Press.
Rainò, P. & Vik, G.-V. (2020). Tulkkausalan tulevaisuudennäkymät. Humanistinen ammattikorkeakoulu julkaisuja 113. Helsinki: Humanistinen ammattikorkeakoulu (Humak). https://urn.fi/URN:NBN:fife2020102185861
Rissanen, T. (1985). *Viittomakielen perusrakenne*. Helsinki: Helsingin yliopiston yleisen kielitieteen laitoksen julkaisuja 12.
Safar, J. & Webster, J. (2014). *Cataloguing endangered sign languages at iSLanDS. iSLanDS's blog: Live reports from the International Institute for Sign Languages and Deaf Studies* (iSLanDS). https://islandscentre.files.wordpress.com/2014/08/report-endangered-sls_070814.pdf
Sallabank, J. (2013). *Attitudes to endangered languages: Identities and policies*. New York: Cambridge University Press.
Salmi, E. & Laakso, M. (2005). *Maahan lämpimään: Suomen viittomakielisten historia*. Helsinki: Kuurojen liitto.
Salonen, J., Kronqvist, A. & Jantunen, T. (2020). The corpus of Finnish Sign Language. In *Proceedings of the Ninth Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives (LREC2020)*. Paris: European Language Resources Association (ELRA), 197–202. https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/SIGN2020book.pdf
Salonen, J., Takkinen, R., Puupponen, A., Nieminen, H., & Pippuri, O. (2016). Creating Corpora of Finland's Sign Languages. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, & J. Mesch (eds.), *Workshop Proceedings: 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining / Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 179-184). Paris: European Language Resources Association (ELRA), 179–184.
Salonen, J., Wainio, T., Kronqvist, A. & Keränen, J. (2019). *Suomen viittomakielten korpusprojektin (CFINSL) annotointiohjeet.* 2. versio. Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto. http://r.jyu.fi/ygR
Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S. & Cormier, K. (2013). Building the British Sign Language corpus. *Language Documentation and Conservation* 7, 136–154.
Viittomakielilaki [Sign Language Act] (359/2015). https://finlex.fi/fi/laki/alkup/2015/20150359 (accessed 24.11.2022)
Takkinen, R., Salonen, J., Puupponen, A., & Nieminen, H. (2020). Miten viittomakielen korpusta luodaan ja mihin sitä tarvitaan? Viittomakielten korpukset ja niiden tehtävät. *Puhe ja kieli* 40 (1), 61–82. https://journal.fi/pk/article/view/95499
The University of Jyväskylä, Sign Language Centre (2018). Finnish Signbank. Available in the Language Bank of Finland (Kielipankki). https://signbank.csc.fi. [Last accessed 24.11.2022].

The University of Jyväskylä, Sign Language Centre (2019). Corpus of Finnish Sign Language [sign language corpus]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2019012321
Wallin, L. & Mesch, J. (2021). *Annoteringskonventioner för teckenspråkstexter. Version 8*. Avdelningen för teckenspråk, Institutionen för lingvistik, Stockholms universitet. https://su.diva-portal.org/smash/get/diva2:1556259/FULLTEXT01.pdf
Wallvik, B. (2005). *Du måste vara döv för att förstå. Finlandssvenska dövas fotspår i historien*. Helsingfors: Finlandssvenska teckenspråkiga r.f.

Ethical Research with Social Media Data: Informed Consent in Large-scale Quantitative Studies

Erwan Moreau[1], Carl Vogel[1] and Kieran Walsh[2]
[1] Trinity College Dublin, [2] National University of Ireland Galway
E-mail: moreaue@tcd.ie

Abstract

Social media have become a common source of research data, offering high volume, high diversity and ease of access. Nevertheless, beyond the basic legal requirements, the ethics of mining social media data is quite complex. In this paper, we briefly review the state of the art recommendations and propose a specific and practical approach through the example of the Virtual-EngAge project, a project in which computational methods are employed for a quantitative study. In the context of designing this project, we analyze the questions of consent and privacy in detail, discussing the limitations of informed consent in particular. Through the perspective of Information Ethics, we advocate for a holistic understanding of the ethics issues related to using social media users' data, as opposed to the standardized ``box-ticking'' approach that informed consent forms may tend to favour. We conclude that explicit consent is not always required, in particular if the outcome of the study is in aggregated form, i.e. in such a way that individual data is not released outside its original context.

Keywords: social media, ethics, consent, privacy

1. Introduction

Many studies, especially in the field of Artificial Intelligence (AI), take for granted that public social media data is available for research purposes and does not require any form of researcher-independent ethics review. From a strictly legal perspective, this assumption may actually be correct in many jurisdictions: assuming that the authors of the content are not minor, that the study does not expose any sensitive information and does not put the authors at risk, the study may simply be exempt of Institutional Review Board (IRB) review (Moreno et al. 2013).

However, from an ethics perspective, the literature clearly calls for caution. Social media users may not fully comprehend the privacy issues which could potentially result from making their data public, and they may not appreciate how their posts could be perceived outside their original context. Somewhat paradoxically, given that Twitter users appear to know their postings will be visible to the general public and accept that re-Tweeting is facilitated by the platform, Williams, Burnap and Sloan (2017) show that 80% of Twitter users have an expectation that their consent would be asked before their content is republished for the purposes of research. At a deeper level, it can even be argued that the traditional model of informed consent is fundamentally biased (O'Connell 2016): due to inherent biases in the researcher-participant relationship, it is argued that the information provided to participants has no or little effect on their decision to participate or not. Thus in the context of online data, O'Connell (2016) considers that the procedure of informed consent is actually not designed to protect the participants, but to protect the researchers and their institution.

The Virtual-EngAge project includes an observational study based on Twitter data, aimed at determining attitudes towards technology among older adults and perceptions held in the community about technological attitudes of the ageing population. The designing stage of the project lead the authors to study the state of the art but also to question some standard approaches to ethics in this kind of study, and eventually to propose a different perspective on the topic.

Thus in this work we briefly review the existing ethics recommendations that have been developed specifically for social media-based research, e.g. (Townsend and Wallace 2016). Consistent with the unanimously adopted position of the Association of Internet Researchers that the guidelines they suggest are not rules (Franzke et al. 2020), we present our approach to the ethics

questions relevant to our project. In this approach, we try to balance the constraints of a small-scale project with these ethics recommendations. In particular we strive to design the study in a way which meaningfully protects peoples' privacy, which gives the participants options with respect to the use of their data, and to the extent possible which anticipates and prevents any form of harm to the participants resulting from the study.

In particular, we argue for strong data availability principles and responsibilities: as long as people make data available for general public consumption (as opposed to privileged consumption, which requires platform membership to inspect) then the data is available for non-harmful research (if people grow roses at the public edge of their front gardens, they have no means of stopping research that depends on counting the publicly visible rosebuds); however, researchers have a responsibility, if requested, to delete underlying data that remains linked to anyone who requests such (an opportunity not available to those whose rosebuds are counted); and people who construct social media data have a responsibility to understand the terms and conditions of social media providers and to be sensitive to public notices of data consumption (that we argue) researchers should provide. That is, as in recent analysis of ethics in active participatory research (Koutsombogera & Vogel 2017), we emphasize that even passive, "involuntary" research participation entails responsibilities, as well. This paper is organized as follows: we present the context and the main ethical issues about mining social media data in Section 2. Then we detail our specific approach in Section 3, and finally propose an in-depth discussion about informed consent and privacy in Section 4.

2. Social Media Data: Ethical Issues

From the ethics perspective, the use of social media data for research purposes is a complex issue. (Olteanu et al. 2019, 21) summarizes it as the difficulty to reconcile two opposite perspectives: "1) social data research is similar to clinical trials and other human experiments in its capacity to harm people, and thus should be regulated as such; and 2) social data research is similar to other computing research, traditionally focused on methods, algorithms and system-building, with minimal direct impact on people."

On the one hand, one can understandably question why researchers should bother with consent and privacy issues beyond legal requirements, given that private actors (starting with the platforms themselves) do not hesitate to monetize private data. In this naive point of view, since the users' data is already exploited to the maximum extent of the law with very little consideration for the users' privacy or even the indirect harm that this can cause them, the use of the same data for research purposes may seem inoffensive or benign. Furthermore, public social media data may also be seen as secondary data:[72] 1 by definition it has already been collected by the platform which stores and publishes it, so it is tempting to assume that the platform is responsible. According to this view, social media data can be used for research purposes without IRB review or through an expedited review process. A large number of studies relying on social media data are published without any form of ethics approval process. This is especially common in disciplines which are not particularly familiar with research based on human participants, like AI, as opposed to social sciences.

On the other hand, researchers' views of ethical use of data should not be automatically benchmarked against the normative standards/values of the industry given differences in goals. There have been a number of works in the literature which specifically study the ethical aspects of using

[72] Secondary data, in a research sense, is typically collected for a specific, or at least a more defined purpose where the scope of potential uses can be argued to be clearer. Where official statistics and data are used, there is less potential to draw out attitudinal inferences or a detailed picture (as in identifiable) of personal circumstances.

social media data in research, as well as recommendations such as those from the Association of Internet Researchers (AoIR) (Franzke et al., 2020). While institutional guidelines vary widely, there is a consensus in the literature that proper ethical considerations should be carefully studied before proceeding with any study based on social media data. Some of the main principles which guide ethics in research were established in the Belmont Report (US National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1978): respect for persons, beneficence, and justice. Under these core principles, researchers have a duty to consider the potential implications of their research on the participants it relies on. There are many ways in which using social media data could directly or indirectly cause harm to the authors of this content themselves or to others. In particular, researchers often underestimate how difficult it is to safeguard the users' anonymity, and consequently their privacy: even when good faith efforts are made at anonymizing the data, it has been shown that simple investigation techniques, often by crossing the dataset with other sources, can uncover the identity of at least some of the participants, if not all (Zimmer, 2010). This can even lead to disclosing a participant's membership of a minority group, potentially making them a target for various forms of discrimination. One should also be careful about the indirect impact the research results or methods can have. For example, a study demonstrates that individuals of low socio-economic status (SES) are more susceptible to some specific disease; the researchers would probably hope that their result will improve the level of healthcare received by this population, but it might in turn cause private insurance companies to limit their access to health insurance, causing the opposite result. In terms of the responsibility of participants to be aware of these issues and possible data uses, there is asymmetric information availability where those creating/posting this data are not aware of the full intended set of uses (particularly given the rapidly changing capacities of systems), and where there can still be argued a duty of care to participants on the part of researchers.

3. Approach

3.1. The Virtual-EngAge Project

The Virtual-EngAge project aims at tackling the challenges of digital exclusion and limited opportunities for engagement and participation for older people. Although these challenges have become very evident during the COVID-19 pandemic, they represent longstanding issues in Irish and other societies. The project is examining how everyday communication devices (such as telephones, and internet enabled devices e.g. smart phones; ipads, etc) could be used by retirement association groups to strengthen their capacity for supporting their members and others to become socially connected, to access and disseminate critical information, and to advocate on key issues for older people. By doing this the project hopes not only to find new short-term strategies for these groups based on the technologies that they have available, but also inform the development of new and usable technologies that are effective in enhancing these sorts of engagement. Currently, ageing related technology is often developed without consultation with older people, and does not reflect their needs, preferences or daily lives.

3.2. Methodology

The Virtual-EngAge project includes an observational study based on social media (primarily Twitter) data aimed at determining attitudes towards technology among older adults and attitudes in their context about the attitudes of older adults towards technology. In this part we sketch the methodology

designed for this study, taking into account the ethical considerations outlined above.

The scope of the study will be determined by a set of target terms, and the social network Application Programming Interface (API) will be used to extract content relevant to the study. Importantly, the goal of the study does not require any individualized data, only global patterns representative of the general attitudes with respect to the topic. While we seek to separate attitudes of older adults from attitudes about older adults, we approach this with a level of granularity that does not require knowing which individual professes which attitude. Naturally the processing, cleaning and interpretation of the data necessarily starts with the raw content posted by the users, which may include personal information (for example the Twitter handle, at least).

Thus we distinguish two stages (described below) in the processing of the data. This is meant to clearly identify the status of the data with respect to privacy, and consequently the required level of safeguarding associated with it.

- The raw data as it appears at collection stage, which potentially contains personal or possibly even sensitive information. This form of data must be treated with extreme care: during the period of retention of raw text data, this will be maintained in password-protected files on secure computer systems. This is in accordance to the appropriate legislation which are underpinned by the Data Protection Acts 1998 and 2003. This entails the following duties:
    - Obtain and process information fairly.
    - Keep data only for specified, explicit and lawful purposes.
    - Use and disclose data only in ways compatible with these purposes.
    - Keep data safe and secure.
    - Keep data accurate, complete and up-to-date.
    - Ensure that data are adequate, relevant and not excessive.
    - Retain data for no longer than necessary.
    - Give a copy of his/her Personal Data to an individual, on request.
- The processed data is a refactored version of the data which does not contain any identifying information. Various techniques will be used to minimize the risk of ulterior personal data collection by third parties:
    - Wherever possible, text content will be represented as “bag of words”, i.e. not in the original word order. This is meant to prevent that an automatic search of a sequence of words would trace back to the online content of the original author. Verbatim quotes of online content will be avoided in publications for the same reason.
    - Specific identifying information in the content, such as persons or locations names, could potentially be used by third parties to retrieve personal data and/or cross it with other datasets. Such information will be removed automatically by filtering out rare words (frequent proper names, e.g. *Ireland*, cannot be used to identify the author).
    - We anticipate that most (if not all) of the published results obtained from social media analysis will represent aggregated trends in attitude among the observed population, as opposed to individual traits. Therefore, the risk of personal data leakage is low.

The research team will establish a process which transforms the raw data into the processed data. This process should be as deterministic as possible, i.e. preferably automated or following clear guidelines when involving human intervention and interpretation. The guidelines will be established progressively from relevant observations made both in the data and externally, with the objective to make the process reproducible. This is needed to ensure that the modifications to the raw data, in particular the removal of some content required by a participant, does not prevent or hinder reaching

the outcome of the process. It is also preferable for replicability purposes,[73] in order to allow a similar experiment to be conducted on a different dataset. This is especially important since the raw data will not be publicly released, thus making reproducibility by other researchers impossible.

Of course, participants retain the right to opt-out of the study and to modify or suppress their content.[74] Only the researchers of the project will have access to the original non-anonymized content (raw data), which will be definitively deleted by the end of the project.

Legally speaking, specific consent is not required because users agreed to the use of their data for various purposes, including academic research, as part of the terms of use of the social media platform. But from the perspective of research ethics, the option to simply rely on the terms of use is questionable. A survey of existing work suggests to follow the ethics recommendations considered the broadest and safest, namely to require participants to sign a consent form. While this option is satisfactory is many cases, it also has significant issues and limitations, detailed in the next section. Thus it was decided for the Virtual-EngAge project to adopt a different approach where consent is not asked but strong measures are taken to protect privacy. In the next section below, we analyse the arguments which led us to this conclusion.

In our view, it is also important from the methodological standpoint that this study is integrated within a broader project, with other strands that capture multiple views, perspectives and lived experiences on the issue. As a mixed-method study, and in overarching terms and in relation to all its methods, it is argued that it benefits from having the considerations and sensitivities of some of the other strands mixed in with decisions around our general approach and design.

4. Discussion about Consent

From a legal perspective, users who post content on social media networks should be aware of the terms and conditions of the platform which hosts their data. In the case of Twitter, users agree to the use of their data for various purposes, including academic research. There is no ambiguity about the legal responsibility of the user; for example, US Courts have confirmed that a person cannot invoke their right to privacy with respect to writings that they post on a social media website, since they made them available to the public by doing so (Moreno et al. 2013).

One principle is that if one accesses data without platform privileges, because the platform and user both make the data public (as through Facebook, perhaps), then the researcher does not have a direct means of contacting individuals, and this distance appears to be appropriate in the context of an observational study. Accessing data with privilege as a member of the platform means that the researcher is more entwined with the prospective research participants, and this seems more complicated, ethically. The responsibility to communicate directly with participants is greater, and they may be tempted to alter their online behaviour thus causing bias in the study.

Nevertheless, the legal framework is often a vague abstraction for many users. In practice, people often do not read the terms and conditions and sometimes do not even have a good understanding of the privacy and security settings provided by social media platforms (Beninger et al., 2014). As (Williams et al., 2017, p. 1153) mentions, “researchers should not assume all users have read and understood terms of service that govern issues such as consent and privacy”. Moreover

[73] We use the ACM terminology: reproducibility refers to redoing an experiment using the same experimental setup, including the same dataset; replicability refers to redoing an experiment using a different experimental setup, for instance a different dataset.(Association for Computing Machinery 2016)

[74] Explicit requests to opt-out are unlikely, since participants would not usually be aware of the study. However a participant could delete their content from public view (e.g. by changing their privacy settings), and this would have the same effect.

(Swirsky, Hoop & Labott 2014, 1) emphasizes that "users may not fully appreciate the privacy risks involved in sharing information, and they may therefore experience an online disinhibition effect". Thus users may feel ashamed or humiliated if their content is taken out of context and scrutinized afterwards. Therefore this could potentially breach the "do no harm" fundamental principle of research ethics.

The major question thus focuses on the extent to which a researcher should protect the social media users when using their data in an experiment, even though the users submitted said data to public scrutiny voluntarily. To phrase the same idea in a somewhat provocative way, is it the responsibility of the researcher to protect the users against their own possible ignorance regarding the service that they choose to use?

4.1. Informed Consent

Traditionally, this problem is answered through a simple consent form: by explicitly asking the participants to consent to the use of their data for a specific and clearly stated research purpose, the researcher can safely assume that the participants have been informed and carry on with their research. While this approach intuitively makes perfect sense and is generally considered satisfactory by ethics review boards, it relies itself on some questionable assumptions.

The principle of informed consent originates from medical research. It was developed as a way to prevent unethical experiments, in the aftermath of some infamous cases of abuse such as the Tuskegee Syphilis Study, from 1932 to 1972. It is established in various international and national legislation, e.g. the Council of Europe's Convention on Human Rights and Biomedicine.

The fact that informed consent is primarily intended for medical studies is often considered as a problem by non-medical researchers, in particular in social sciences, for several reasons. First, it is clear that the risks are of a different nature when the participants undergo some medical procedure versus when their social media data is analyzed. At a deeper level, "many argue that informed consent protocols reproduce a dominant medical model and a rigid view of power in clinical and non-clinical research.[75] Universal standards never uncover the material and cultural inequities of research itself (i.e., North–South funding inequities), the problematic assumptions underlying research studies (Western colonial epistemes), and the institutional and organizational practices that configure the researcher and participant role (university vs. community)" (O'Connell 2016, 73). There is a fundamental imbalance of power between the researcher, i.e. an authority figure, and the laypersons asked to consent to some apparently complex research on their data. This bias can cause people to sign a consent form without reading or understanding it, making their consent ethically meaningless, but legally valid. In fact, it has been shown that "providing (too much) information to the research subject can occasionally lead to the opposite effect of what the informed consent aims at; excess of information can leave the concerned party unable to make a (truly) informed choice after all." (Christen et al. 2016, 209).

From this point of view, it can reasonably be argued that informed consent forms are designed to primarily protect the researchers and their institutions, not the participants. Practically, these are used as a legal contract signed by a supposedly rational person, releasing the researchers (and their institutions) from any further scrutiny. As long as the participant signed the form, it is assumed that they have a clear understanding of the goals and risks of the project, even though the information they

[75] Of course, in social science research, power imbalances may also exist, as for example in situations in which the researcher conceives of an issue as a problem, but where prospective participants do not all agree that the issue constitutes a problem.

are provided with has actually very little effect on their decision-making process (O'Connell 2016). "The normative top-down expression of power ... [imposes] ... a singular standard for consent that is based on the idea that the researcher always has more power and no risk compared to the participant" (O'Connell 2016, 74).

In the context of social media data, there is an irony in asking users to give their consent: this is usually done through a form which explains the research in fairly technical language and explains to the users their rights in legal terms, essentially reproducing the same kind of bias found in the terms of service of the platform. Many users perceive this document as long and full of obscure jargon, and end up not better informed or truly consenting whether they agree or not. Since the motivation for asking their consent was precisely to make sure that they agree assuming that they might not have read or understood the terms of use of the platform, it seems misguided to assume that they would this time truly read and understand the consent form. As a consequence this process is ethically meaningless: if one assumes that users are reasonably careful and rational, their consent to the terms of usage of the platform is sufficient for using their data. If it is assumed that users do not truly understand these, it is extremely questionable to expect them to better understand the research consent form.

Naturally, there are certainly cases where the informed consent plays an important role and duly protects participants. But the standardisation of informed consent processes is sometimes akin to an industrial automation process, i.e where a task previously relying on human expertise becomes "de-humanized" for the sake of efficiency, especially within social media platforms: in this "simplified" approach, the notion of consent is codified and formatted in a way which facilitates a "tick-box approach" to ethics, where instead of considering the diversity of individuals and the various potential difficulties or questions they might raise, their understanding and consent is extremely simplified into a polar interrogative: did they sign the consent form? The Facebook-Cambridge Analytica data scandal[76] started with a study which duly obtained the "informed consent" of many Facebook users, thus had at least the appearance of ethical legitimacy. The lack of any check or monitoring by Facebook and the dishonest behaviour of Cambridge Analytica show why consent forms on their own do not suffice to make a study ethically valid: real ethics is not a matter of formal "box ticking", it is a continuous process which requires efforts by the researchers, their institutions and their research communities towards making sure that every step is done in accordance with ethical principles.

Of course, practices widely vary among different fields, institutions and researchers: not every researcher or IRB follows a simplistic approach to ethics, nor is it necessarily prevalent. While this shift in thinking around the nuances and challenges of consent has become evident within certain research fields and institutional research review boards, for example by developing active and iterative consent and assent processes, it can be argued that research concerning social media analytics has remained narrow in its understanding and practices regarding consent.

4.2. Privacy

Asking the participants their consent is meant to fulfill several objectives: their agreement validates the fact that they are aware of the risks for themselves, and of the rights and protections that they are offered. It also confirms their support, or at least their absence of objection, to the goal of the study. This can be important in case the study involves a sensitive and/or controversial topic, such as abortion rights or the rehabilitation of former convicts.[77] When collecting social media data, the main

[76]https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal – last verified May 2022.
[77] In the case of the Virtual-EngAge project, it is reasonably safe to assume the goal is not sensitive or controversial.

risk to participants and thus the main ethical issue that informed consent aims to address is about their right to privacy, and the harm which can result if their privacy is breached.

4.3. Information Ethics

In this analysis of the question of privacy, we try to apply and follow the principles of Information Ethics (IE) (Floridi 1999). IE is proposed as a macro-ethic which does not only provide a solid basis for Computer Ethics (the field of ethics applied specifically to issues in the domain of new technologies), but also offers an original perspective by making information the main focus of any ethics question. Here information should be understood in the broadest sense: "any entity is a consistent packet of information, that is an item that contains no contradiction in itself and can be named or denoted in an information process." (Floridi 1999, 43). In this perspective, the integrity of the infosphere (the information environment) should not be damaged, should be preserved, enriched and nurtured. The author expresses this by proposing four laws (ordered by increasing moral value) which determine whether an action is moral or not. Entropy in the infosphere ought: (1) not to be caused; (2) to be prevented; (3) to be removed. (4) Information welfare ought to be promoted by extending, improving and enriching the infosphere. On this construction, as entities, humans constitute information, as do rosebuds, but facts also supply information, while falsehoods do not. Our view of information ethics is that it provides an imperative to develop information, to study what may be learned from data that is visible in public without privileged access, provided this de-links data from individuals from the analysis and reporting (despite individuals having themselves created a link to the primary data).

4.4. A different perspective on privacy

Privacy is a complex question. (Coll 2014, 1250) argues that the concept of privacy has been "reshaped by and in favour of informational capitalism, notably by being over-individuated through the self-determination principle". As a result, "privacy becomes only about data and remains the right and responsibility of every individual instead of a collective value." (O'Connell 2016, 81). In this perspective, (O'Connell 2016, 82) also argues that "concerns about big data sets become a question of data protection, not a question about the ethics of the research question. As a less direct form of data collection, issues of harm and confidentiality appear less critical or are viewed as being already in the public domain." This is a serious issue, because restricting the concept of privacy masks the fact that ultimately the risk is about harm, and increasingly the risk of causing distress within that. Thus despite an individual's privacy being protected, if what the individual said or their actions is framed in a way that causes them distress, this is just as problematic.

This concept of data protection stems from the view of privacy as an individual's right to decide whether they want to retain or release information about themselves. This view assimilates personal data as a property of the individual, and naturally the individual is entitled to do as they see fit with their property, similarly to their physical properties. IE offers a significantly different interpretation: it postulates that their information does not only belong to the individual, the individual is the sum of all their information. Thus any privacy "intrusion is disruptive not just because it breaks the atmosphere of the environment, but because any information about ourselves is an integral part of ourselves, and whoever owns it possesses a piece of ourselves, and thus undermines our uniqueness and our autonomy from the world. There is information that everyone has about us, but this is only our public side, the worn side of our self, and the price we need to pay to society to be recognised as

its members" (Floridi 1999, 53).

In the context of collecting social media data, we propose the following interpretation: social media users post content[78] voluntarily on a platform. Their action takes place in a specific context, i.e. time and environment (social circle, chances that strangers would see the content, etc.), which defines the boundaries in which the user intends to broadcast this information (whether they are fully aware of these boundaries or not). As a consequence, and given the current inefficiencies and flaws in the process of consent declarations from users, we argue that collecting data without further consultation is ethically acceptable as long as it does not modify the boundaries defined (purposefully or not) by the individual, and as long as a macro-ethic applies regarding data aggregation and dissemination. This implies that their data should not be broadcasted outside the original context. For example, verbatim quotes in a research article should be avoided because they make the identification of the participant easier, since entering the quote on a search engine is usually sufficient to find the original post. Additionally identification should be prevented not only by a random stranger, but also by people belonging to the social circles of the author, e.g. members of their school, work environment, neighbourhood. Instead, all the results should be aggregated in a way such that the original individual content is indiscernible. For example, (Williams et al. 2017, 1158) suggests that "quantitative analysis of Twitter data that presents findings in aggregate form (such as tables of regression results, topic clusters in word clouds and anonymised network visualisations) is one way to support ethical research without the need for informed consent."

4.5. The context matters

Naturally the participants and their potential level of vulnerability and marginality are an essential component of the ethical design of a study. For example, the Virtual-EngAge project is focused on older adults and their attitudes towards technology in Ireland. The general arguments that we put forward in this article apply, but additionally there could be some prejudice in the population around this topic, and a forum such as Twitter is prompt to mock or even insult people for their mistakes. This makes protecting the participants' privacy (in the sense described above) a priority of our ethics design.

The objectives of the study also matter, in particular the type of information collected as well as the audience susceptible of having access to it. For example, safeguarding the participants' privacy requires a stronger approach if the dissemination plan involves media outreach and political organizations than if it plans only scientific articles in a few specialized journal.

It is important to emphasize that the authors do not support any approach to the exclusion of all others. On the contrary, like many others we strongly encourage taking the specific context of the study into account in the ethical design, as opposed to adopting any predetermined solution. In particular we acknowledge that multiple other factors can also be taken into account, even though we did not address them specifically in this paper: cultural systems inform ethical values, and the evolving international standards should not dismiss regional interpretations for example.

5. Conclusion

We have argued that the imperative to create knowledge suggests that learning generalizations by

[78] In the IE view, any content that a person posts is personal information: even if it does not contain anything about the author themselves, the simple act of posting is itself an information about the individual, and therefore a part of the individual.

aggregating data voluntarily made available to the general public by individuals entails that it is ethical to so study such data. For millennia the imperative to create knowledge has been balanced by fear that knowledge can be dangerous and that there are certain things which should not be known. But in these current circumstances, it falls to the researcher to ensure that a macro-informatic ethic is applied to ensure this aggregation is a sufficient abstraction to ensure privacy and freedom from harm and distress. Further work is required within the field to guide and perhaps regulate this aggregation to support researchers in this endeavour.

We argue that provided one abstracts away from the individuals who create data, and eschew identifying individuals in reporting data and generalizations, where people have voluntarily made data visible to anyone in the public who lacks privileged access, it is ethical to study that data without additional consultation. In most cases, it is appropriate for such researchers to provide a similarly unfettered declarations of their research. More sophisticated platforms for such declarations may emerge in the future, and development and innovation work in this regard is certainly required. However, regardless of whether they do or do not, there is a critical need to address urgent ethical questions regarding how knowledge gained through these spheres is applied and disseminated. It is useful to know whether rosebuds are opening earlier each year; it is wrong to use this knowledge to impune individual gardeners.

6. Acknowledgements

This research is supported by the Irish Research Council through its funding to the Virtual-EngAge project (COALESCE/2021/63).

References

Association for Computing Machinery (2016). *Artifact review and badging*. https://www.acm.org/publications/policies/artifact-review-badging

Beninger K., Fry, A., Jago, N. Lepps, H., Nass, L. & Silvester, H. (2014). Research using social media; users' views. *NatCen Social Research*, 1–40.

Christen, M., Domingo-Ferrer, J., Draganski, B., Spranger, T., Walter, H. (2016). On the Compatibility of Big Data Driven Research and Informed Consent: The Example of the Human Brain Project. In Mittelstadt, B., Floridi, L. (eds), *The Ethics of Biomedical Big Data*. Law, Governance and Technology Series 29. Springer, 199–218.

Coll, S. (2014). Power, knowledge, and the subjects of privacy: understanding privacy as the ally of surveillance. *Information, Communication & Society* 17 (10), 1250–1263.

Floridi, L. (1999). Information ethics: On the philosophical foundation of computer ethics. *Ethics and Information Technology* 1, 37–56.

Franzke, A. S., Bechmann, A., Zimmer, M., Ess, C. & the Association of Internet Researchers (2020). *Internet research: Ethical guidelines* 3.0. https://aoir.org/reports/ethics3.pdf.

Koutsombogera, M. & Vogel, C. (2017). Ethical responsibilities of researchers and participants in the development of multimodal interaction corpora. In P. Baranyi, A. Esposito, P. Földesi &T. Mihálydeák (eds.), *8th IEEE International Conference on Cognitive Infocommunications* (CogInfoCom 2017). IEEE, 277–282.

Moreno, M. A., Goniu, N., Moreno, P. S. & Diekema, D. (2013). Ethics of social media research: common concerns and practical considerations. *Cyberpsychology, behavior, and social networking* 16 (9), 708–713.

O'Connell, A. (2016). My entire life is online: Informed consent, big data, and decolonial knowledge. *Intersectionalities: A Global Journal of Social Work Analysis, Research, Polity, and Practice* 5 (1), 68–93.

Olteanu, A., Castillo, C., Diaz, F. & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (13).

Swirsky, E. S., Hoop, J. G. & Labott, S. (2014). Using social media in research: new ethics for a new meme? *The American Journal of Bioethics* 14 (10), 60–61.

Townsend, L. & Wallace, C. (2016). *Social media research: A guide to ethics.* University of Aberdeen 1 (16).

US National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1978). *The Belmont report: ethical principles and guidelines for the protection of human subjects of research* 2. Department of Health, Education, and Welfare.

Williams, M. L., Burnap, P. & Sloan, L. (2017). Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology* 51 (6), 1149–1168.

Zimmer, M. (2010). "But the data is already public": on the ethics of research in facebook. *Ethics and information technology* 12 (4), 313–325.

Developing Automated Feedback on Spoken Performance: Exploring the Functioning of Five Analytic Rating Scales Using Many-facet Rasch Measurement

Anna von Zansen, Ari Huhta
University of Helsinki, University of Jyväskylä
E-mail: anna.vonzansen@helsinki.fi

Abstract

In this study, we used the Many-facet Rasch measurement (MFRM) to explore the quality of ratings as well as the functioning of five analytic rating scales developed for automated assessment of L2 speech. This study is part of a multidisciplinary research project that develops automatic speech recognition (ASR), automated scoring and automated feedback for L2 Finnish and Swedish. The data include the analytic ratings (task completion, fluency, pronunciation, range, accuracy) gathered from human raters (n=14) who assessed L2 Finnish learners' (n=64) speech samples using Moodle. The four-facet Rasch analysis showed that the raters performed and the rating scales functioned well, although task completion seems to be more challenging to apply consistently than the other criteria. Moreover, it proved to be more difficult to receive a certain score on some dimensions, namely fluency and range, than others. The study has implications for score reporting. We demonstrated that a) the different analytical rating scales have somewhat different structure, b) scores do not advance with equal intervals and c) a certain score on a certain dimension might require a bigger leap forward in ability than on other dimensions. The results will be used for designing encouraging and accurate automated feedback to L2 Finnish and Swedish learners.

Keywords: automated feedback, language assessment, rating scales, oral skills

1. Introduction

Automated speech processing technology has improved and become more popular in everyday life contexts. Also, the technologies for automated assessment of speaking skills have made considerable progress in recent years. Automated language assessment has many advantages: not only can it save time and money, but it can also standardize the scoring process. However, automated systems still have many limitations, for example, regarding construct coverage. Therefore, a hybrid approach that combines human and automated scoring is likely to be the most feasible solution (see Evanini & Zechner, 2020, 3–4; Xu et al. 2020).

As Gu & Davis (2020, 159) point out, automated speech processing technology can be used to provide immediate and individualized diagnostic feedback to L2 learners, regardless of time and place. Moreover, automated systems can give such feedback instantly (Zhang et al. 2020, 21). Automated feedback technologies are emerging also in language learning contexts, where tutoring systems can provide immediate and specific feedback or instruction to the learner (Golonka et al. 2014, 73). However, most of the automated language learning tools deal with written language and grammar (for a review of educational feedback systems see Deeva et al. 2021).

Turning to L2 speaking, most ASR-based software for training speaking are limited to computer-assisted pronunciation training (Golonka et al. 2014, 81), although de Vries et al. (2015) present an ASR-based system developed for practicing word order in Dutch. Many automated speech training systems such as EduSpeak, NativeAccent, English Discoveries and Duolingo provide feedback mainly on pronunciation (Gu & Davis 2020, 159–160). Nevertheless, some automated systems that provide feedback on spontaneous speech exist but the tools are often aimed only for L2 English learners. For example, in the context of TOEFL Practice Online Test, Gu & Davis (2020) describe the development of automated feedback on seven features related to speaking, whereas Xu et al. (2020) present validity argument for the Linguaskill Speaking Test which combines auto-scoring and human rating to produce a CEFR grade to the L2 English learner.

This study is part of the DigiTala research project (2019–2023) which develops an automated tool for assessing L2 Finnish and L2 Swedish learners' oral skills (see Kautonen & von Zansen 2020). In this multidisciplinary project, experts of pedagogy, technology and phonetics develop automatic speech recognition, automated scoring and automated feedback (see Evanini & Zechner 2020) for assessing L2 Finnish and Swedish learners' oral skills.

The research project has two goals: 1) to pave way for implementing a speaking section to the language tests of the Finnish Matriculation Examination (Vaarala et al. 2021) and 2) to develop an online tool for self-regulated learning purposes. The study reported here relates mostly to the second goal. The aim of the automated diagnostic feedback is to help both independent learners and learners with access to teacher support to develop their speaking skills by providing information about the strengths and weaknesses in their performance.

In this study, we use Many-facet Rasch Measurement (MFRM, see McNamara et al. 2019; Boone et al. 2014; for a review of Rasch measurement in language assessment see Aryadoust, Ng & Sayama 2021) to explore the functioning of the analytic rating scales which are used for designing diagnostic feedback on speech performances.

1.1. The Moodle plugin

The project designed a Moodle plugin (von Zansen et al. 2022) that records L2 learners' responses to a speaking task and displays automatically rated scores to the learner. Currently, the task types include read-aloud and spontaneous speech (up to 3 minutes). When the system receives a speech sample, it uses automatic speech recognition to produce a transcript of the sample. Then the system produces automatic scores on selected dimensions of speech and finally shows the results to the learner.

The Moodle plugin's (von Zansen et al. 2022) frontend is described in a user manual that presents the Moodle plugin in detail (Alanen et al. 2022). Moreover, a short video is available on the Github page (von Zansen et al. 2022) and a screenshot of the learner's report page is available in Appendix 1.

For the backend, we have trained automatic assessment systems using Finnish and Swedish learners' speech samples that were rated by human raters. We follow a feature-based approach, which enables the production of feedback on different dimensions of speech. However, we are also exploring whether better results could be achieved by applying deep learning methods, "the black box approach" (Al-Ghezi et al. forthcoming).

For read-aloud samples, the system produces scores for fluency and pronunciation while also showing the transcript of the sample to the learner and pointing pronunciation errors. For spontaneous samples, the system provides more detailed feedback: a transcript of the sample combined with analytic scores on fluency (e.g. breaks and repetitions on a 0–4 scale), pronunciation (control of sound and prosodic features on a 0–4 scale), task completion (does the speaker answer the question on a 0–3 scale) and range (extent of vocabulary, structures and expressions on a 0–3 scale) as well as an estimation of the proficiency level (from below A1 to C2 on the Common European Framework of Reference scale, see Finnish National Agency for Education 2003; Council of Europe 2001). The analytic scales include dimensions that human raters are familiar with and hat can be measured automatically (see Kautonen & von Zansen 2020 and section 2 for scale development).

In addition to the automated scoring, teachers have the possibility to comment on the scores produced by the machine. Finally, teachers or researchers can export the learners' speech samples and their scores. The rating data together with the speech samples are important for us in the future when we evaluate the reliability of the automated system (see also Evanini & Zechner 2020, 13).

1.2. Quality of ratings

In this study, we investigate the "quality of ratings", which refers to (a) raters' performance and (b) functioning of the rating scales.

Regarding rater performance, the Facets programme provides information on raters' relative severity, that is, how severely (or leniently) they rate compared to the other raters in the sample (McNamara et al. 2019, 108). Rater severity is not necessarily a significant concern in this project, since we use fair averages to train the automatic scoring system. Fair averages produced by Facets are scores adjusted for rater severity / leniency and they are, thus, more accurate indicators of learner ability than regular (raw) averages calculated across the ratings given to a particular speech sample. Second, Facets produces rater fit statistics, which are informative of rater consistency (McNamara et al. 2019, 109). In this study, we use the range 0.5–1.5 for acceptable fit statistics recommended by Linacre (2002a). High mean-square values (above 1.5) indicate misfit meaning that the rater performs inconsistently. Low mean-square values (below 0.5) indicate that a rater overfits the model which means that the rater shows less variation than was expected, possibly due to halo or central tendency effects (see McNamara et al. 2019, 109). In general, very inconsistent raters (as indicated by above 1.5 mean-square values) degrade the dependability of the rating data. However, also extremely severe or lenient raters are problematic since Facets can adjust the fair average score only up to a point – such extreme cases need to be spotted by visually inspecting Facets output and decisions need to be made whether to remove them from the data).

Rating scale functioning is the second focus area of this study, as we plan to use the rating scales as a starting point when providing automated feedback to the learners. Linacre (2002b) recommends following guidelines[79] for optimizing the functioning of a rating scale. For stable and precise estimates, each score category should have over ten observations (guideline 1). For optimal step calibration, the observations should be regularly distributed across the score categories (guideline 2). Furthermore, average measures should advance monotonically (guideline 3), in other words, higher score observations produce higher measures. In addition, outlier-sensitive MNSQs should be less than 2.0 (guideline 4) since score categories with larger Outfit MNSQs indicate too much randomness ("noise") and are therefore not useful for the measurement (see also Linacre, 2002a). According to the guideline 5, step calibrations must advance, that is, high measures are observed in the highest categories and vice versa. Disordering of step calibration may occur if the construct (speaking ability) is not well defined, or a score category reflects too narrow part of it. Finally, step difficulties (Rasch Andrich thresholds) should advance by at least 1.4 logits (guideline 6), yet less than by 5.0 logits (guideline 7). These guidelines are helpful when evaluating the functioning of rating scales. Sometimes scale revision such as combining neighbouring categories, might be needed, if raters cannot distinguish between such categories (see Linacre 2002b; McNamara et al. 2019, 70–78.)

1.3. Ongoing research and research questions of this study

To develop automated assessment of L2 learners' oral skills, we have followed the stages described in Figure 1. First, we analyzed human ratings after receiving and transcribing the speech samples from L2 Finnish learners. However, these analyses served a different purpose, that is, converting the ordinal rating scale data to linear measures (by using Facets analysis) in order to receive a fairer and more accurate score for each speech sample (see Boone et al. 2014). After these analyses, various machine learning methods are applied to the speech samples and their transcriptions in order to predict

[79] Guidelines renumbered 1–7; guideline 6 (Linacre 2002b) omitted from this study due to the complexity of the analysis

the human ratings. Emerging results (Al-Ghezi et al. forthcoming) suggest that the automated system could predict most of the analytical ratings statistically significantly. The prediction was best for the fluency ratings (Spearman correlation 0.47 for Finnish and 0.23 for Swedish) followed by range (0.28 for Finnish and 0.20 for Swedish) and accuracy (0.22 for Finnish and 0.18 for Swedish). For pronunciation, the correlation between human and machine ratings were significant for Swedish (0.17) but not for Finnish.

In addition to the rating data, we plan to take stakeholders' perceptions (see von Zansen et al. accepted; von Zansen, Sneck & Hilden accepted a; b) into account when designing automated feedback. Moreover, we are interviewing learners and teachers in order to investigate the usefulness and understandability of the automated feedback (von Zansen & Heijala forthcoming).

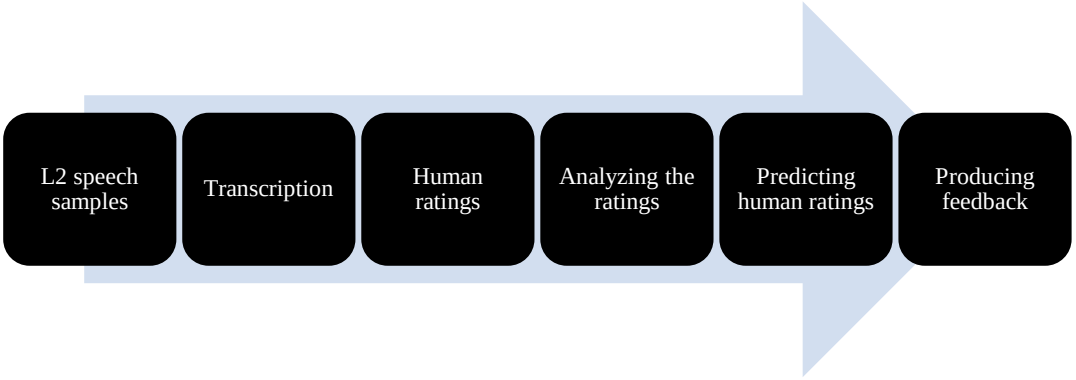


Figure 1: Stages of development

However, we have not yet investigated in detail how the raters performed nor how the scales functioned. Therefore, to address the research gap and to provide evidence for the validity of the human ratings that are important for the overall validity of the automated system, the study seeks answers to two main research questions (RQ): 1. What was the quality (i.e., consistency and agreement) of the ratings across the different analytic scales? 2. How did each analytic rating scale function as a scale?

In other words, we focus on the fourth and sixth stages presented in Figure 1. Results of the analyses support mainly carrying out the last stage (see Figure 1), since the analytic rating scales can be used as part of the automated feedback to learners and as a starting point for developing even more fine-grained feedback on a range of specific features of speech. The results of this explorative study serve proof-of-concept purposes.

2. Methods

The data of this study include ratings gathered from human raters (n=14) during the third rating round organized by the project in June 2021. Speech samples were rated by using a holistic (below A1–C2) and five analytic (task completion, fluency, pronunciation, range, accuracy) rating scales (see von Zansen 2022a) using Moodle. In Moodle, the raters listened to one sample at a time and provided both the holistic and analytic scores in the same window. We used a partially overlapping rating design where some of the samples (n=913) were systematically routed for two or multiple raters to rate while some were rated by a single rater. This way we saved human resources and were still able to investigate and compare all the ratings simultaneously with Facets (Linacre 2021). As a result, in addition to investigating the quality of ratings (RQ1) and scales (RQ2), by using Facets, we obtained fairer scores for training the automatic scoring system (see sections 1.2 and 1.3). During this rating

round we collected over 7500 ratings in total, of which 1030 were holistic scores. For details concerning rater training and instructions see von Zansen, Sneck and Hilden (accepted b).

In scale development, we used the level descriptors of the previous National Core Curriculum (Finnish National Agency for Education 2003) as a starting point for several reasons (see also Kautonen & von Zansen 2020). First, as they came from the National Curriculum and as they are local applications of the Common European Framework (Council of Europe 2001) descriptors, the scale is well-known both nationally and internationally. Secondly, this allowed us to address the first goal of the research project, that is, enabling implementing a speaking section to the language tests of the Finnish Matriculation Examination which aims to measure the outcomes of the level of education regulated by the above mentioned National Curriculum (see section 1). Third, the chosen descriptors suit assessment purposes in general as they describe learner skills in sufficient detail. Thus, their detailed, analytical nature makes them applicable to be used in automated scoring and feedback.

The rated Finnish language samples were collected during spring 2021 from upper secondary school students (n=64) using speaking tests (von Zansen 2022b, 2022c) targeting B1 and B2 levels of the Common European Framework of Reference (CEFR, Council of Europe 2001). Altogether, the tests consisted of eight different tasks and 26 subtasks (von Zansen 2022b, 2022c). Since the data collection took place during the COVID-19 pandemic, data from both raters and L2 Finnish speakers were collected using Moodle and Zoom (see Al-Ghezi et al. forthcoming; von Zansen, Sneck & Hilden accepted a).

To explore the functioning of the analytic rating scales, the ratings were analyzed using Many-facet Rasch Measurement (MFRM, see McNamara et al. 2019) using Facets version 3.83.5 (Linacre 2021). In this four-facet Rasch analysis we included 1) learner's speaking ability (64 speakers), 2) task difficulty (26 tasks), 3) rater severity (14 raters), and 4) difficulty of the criteria (five analytic criteria).

We used a partial credit model to the fourth facet (see McNamara et al. 2019, 115–116), which enables modelling each of the analytic criteria to have its own scale structure. This yields more information about the scales than the rating scale model that assumes all the scales to have the same structure (McNamara et al. 2019, 115–116).

3. Results

To give an overview of the data, we first present the calibration of the four facets as Figure 2. After that, we present results for RQ1 and RQ2. More comprehensive results of the Facets analysis can be found in Appendix 2.

Figure 2 (available also as Table 6.0 in Appendix 2) shows the Wright map, that is, the location of the elements in each facet in relation to the other facets. The measurement scale ("Measr") is an interval scale that ranges from -2 to +4 logits in this analysis and provides a common yardstick against which all the facets and all their elements can be compared. We allowed the first facet (learners' speaking ability) to float while other facets were anchored at zero.

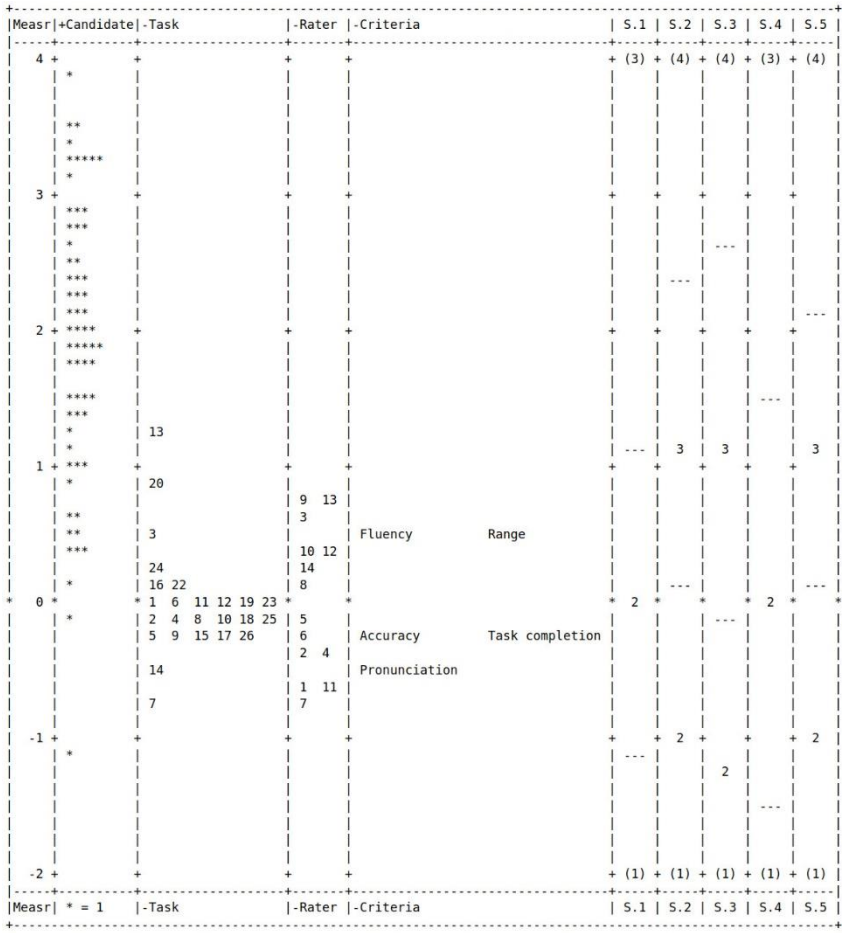


Figure 2: Wright map

In the second “Candidate” column, each star represents a test-taker showing that the learners are spread over five logits meaning that they differ considerably in their speaking ability as measured by the five analytical dimensions analysed here. Learners with higher speaking ability are at the higher end of the logit scale. The next “Task” column shows the tasks organized according to their difficulty. Based on the raters’ analytic ratings the test appears to have been quite easy as there are no tasks matching the best speakers. The tasks are spread over two logits; task 13 being the hardest and task 7 the easiest. The middle “Rater” column arranges the raters according to their severity raters 9 and 13 being the harshest and rater 7 the most lenient.

The “Criteria” column shows the relative difficulty of the five analytic criteria. We see that fluency and range are harder than accuracy and task completion and pronunciation, which is nearly one logit below the hardest criteria. In other words, it is harder for learners to receive a score on fluency or range compared to the other criteria. Furthermore, fluency and range as well as accuracy and task completion are similar in terms of difficulty.

On the right side of the Figure 2, we find each rating scale criterion (S.1 Task completion, S.2 Fluency, S.3 Pronunciation, S.4 Range, S.5 Accuracy) having a separate column. In the brackets, we see the highest and lowest scale levels, for example, the scale for task completion is 1–3 while fluency scale is 1–4. The horizontal lines in the scale columns show the points at which a learner at that logit level would score a half score (Rasch-half-point thresholds, see McNamara et al. 2019, 100). For example, speakers slightly above the logit value of 1 will likely receive a score 2.5 on task completion, 3 on fluency, 3 on pronunciation, 2 on range and 3 on accuracy.

After giving an overview of the data, we now present results regarding the quality of the ratings across the different analytic scales (RQ1). Table 7.3.1 in the Appendix 2 shows details on how the raters performed. Rater IDs are on the right column. "Total Count" shows the number of ratings performed by each rater. We notice that raters 1–4, who were researchers of the project, have provided fewer ratings than raters 5–14 recruited by the project for this rating round. The "Measure" shows raters' severity on the logit scale, which differs by 1.53 logits (rater 13 being the most severe while rater 7 being most lenient). Finally, the "Model S.E." column tells that the estimation of the rater measures is fairly precise especially for raters who provided more ratings while standard error for raters 1–4 is somewhat larger (.13–.15). The fit statistics (see columns "Infit MnSq" and "Outfit MnSq") indicate that all raters fit the model well (Infit MNSQs range .81–1.36, see Linacre 2002a). The reliability of the rater separation index (.96) indicates that the raters are a heterogeneous group. The last row of the Table 7.3.1 in the Appendix 2 shows that the inter-rater agreement was 57.8%, and that the raters agreed more than was expected by the model (51.6%).

Next, we look at findings relating to the RQ2, which deals with the functioning of the rating scales. Table 7.4.1 in the Appendix 2 shows details on the analytic scales used in this rating round. As mentioned earlier, it is more difficult to receive a score on Fluency (.51) and Range (.47) than it is on Accuracy (-.20), Task completion (-.28) or Pronunciation (-.51, see "Measure" column). Nevertheless, the fit statistics indicate that all analytic criteria are within the acceptable range (Infit MNSQs .91–1.20). However, Task completion has quite high MNSQs (Infit MNSQ 1.20, Outfit MNSQ 1.43) and seems to be more challenging to apply consistently than the other criteria. This finding is supported also by the list of unexpected responses (see Table 4.1 in the end of Appendix 2), where the majority of problematic ratings relate to Task completion. Finally, the separation ratio ("Separation" 5.57) and the separation index ("Strata" 7.76) indicate that the analytic dimensions differ in difficulty. The high criteria separation index ("Reliability" .98) shows that the test is measuring different dimensions of speaking rather than speaking as one unitary dimension.

We investigated the functioning of the rating scale following Linacre's guidelines (2002b). Results of the Facets analysis regarding this section can be found in Tables 8.1–8.5, see Appendix 2. Firstly, we noticed that the guideline 1 did not hold for two of the 4-pointed scales. Namely, for Pronunciation, score 1 was given only 6 times and for Accuracy, score 1 was given only 9 times. This might lead to unstable step calibration (Linacre 2002b) for those particular score levels. Second, with regard to the guideline 2, we noticed that the observations were not regularly distributed across the score categories (guideline 2). In general, lower scores (especially 1) were again given less frequently.

Third, we noticed that the average measures advanced (guideline 3). The average measures were also close to the expected values except of Task completion score category 1 (average .98, expected .60) and Pronunciation score category 1 (average 1.28, expected .45). Fourth, the guideline 4 did hold since all the outlier-sensitive MNSQs were less than 2.0. However, score 1 both in the Pronunciation (Outfit MNSQ 1.5) and Task completion (Outfit MNSQ 1.7) scales seems to have more noise that was expected. Fifth, we investigated the probability characteristic curves, and noticed that the score categories appear as a range of hills, indicating that guideline 5 holds. However, we observed one average measure being disordered (score category 2 for Pronunciation, average measure 1.05*, see Appendix 2, Table 8.3), presumably because there were only six observations for the lowest category one, even though it is also possible that the definitions of categories one and two are not clear enough. Sixth, investigation of the Rasch Andrich thresholds showed that guidelines 6 and 7 hold: the step difficulties advanced at least by 1.4 logits but less than 5.0 logits.

4. Discussion

Automated feedback systems are becoming common also in language assessment (Deeva et al. 2021), yet most of the tutoring systems (see Golonka et al. 2014) focus on written language and grammar or target a narrow aspect of speaking, such as pronunciation or word order (de Vries et al. 2015). This study extended previous research by exploring two aspects of automated feedback systems, namely rater performance (RQ1) and the functioning of several analytic rating scales (RQ2) in the context of developing an automated speech training system for L2 Swedish and Finnish learners. Unlike many existing systems, the Moodle-based tool (von Zansen et al. 2022) developed by the research project (Kautonen & von Zansen 2020) provides automated diagnostic feedback on learners' spontaneous speech performances (see section 1.1). To the best of our knowledge, this tool can be compared only with two systems targeted for L2 English learners. First, Gu & Davis (2020) have used a feature-based approach to develop automated feedback to L2 English speakers. Second, in addition to automated analytic feedback, we aim to provide an estimation of L2 speaker's CEFR level (Council of Europe 2001), which is in line with the work of Xu et al. (2020).

The experiment provided new insights into the five analytic rating scales that are, in addition to human scoring (hybrid approach, see Evanini & Zechner, 2020; Xu et al. 2020), also used for designing automated scoring and feedback. The methodological choices of the study proved to be useful for exploring the quality of ratings (RQ1) and the functioning of the five analytic scales developed by the project (RQ2). MFRM has many advantages compared to the more traditional approaches that focus on raw scores or compare pairs of raters. These include converting ordinal rating data into linear measures, creating rating designs that save resources, taking rater severity and task difficulty into account in order to compute fairer scores to the learner (see for example Boone et al. 2014; McNamara et al. 2019; Aryadoust, Ng & Sayama 2021 for a review of Rasch measurement in language assessment).

In this study, MFRM enabled investigating how the raters performed (RQ1) and the step structure of each rating scale criterion (RQ2). Regarding the quality of the ratings (RQ1), results of this study indicate that the raters performed well. The inter-rater agreement (57.8%) exceeded what was expected by the model. Furthermore, we did not observe raters misfitting. The results of the RQ1 indicate that the overall reliability of the raters was good. Recruitment of the raters had been successful, and we had provided sufficient training and instructions to the raters. Turning to the functioning of five analytic ratings scales (RQ2), we noticed that the rating scales functioned reasonably well, although task completion seems to be more challenging to apply consistently for the raters. It appears to measure a somewhat different aspect of speaking than the other scales. However, this is not surprising when we think about the content-relatedness of the concept "Task completion" compared to the more linguistic concepts "Fluency", "Pronunciation", "Range" and "Accuracy". A reasonable conclusion from this finding is that if Task completion is to be used as part of the automated feedback, it needs to be defined more clearly. It is possible, for example, that the meaning of task completion may differ somewhat depending on the particular task or task type, and, thus, ideally, different tasks may require slightly different scales for Task completion.

Ultimately, possible threats regarding the functioning of the rating scale might produce imprecise estimates, which in turn can lead to unfair decisions and conclusions. The results of the RQ2 suggest that some scale revisions might be needed. However, in the case of this study, we think that some of the observed problems result from the small sample of speakers (n=64), which is one limitation of this study. Moreover, the tasks targeted only B1 and B2 level speakers. Due to the lack of A-level speakers, the results cannot confirm whether raters would use the lower end of the scales

reliably enough. More research is needed to verify whether score 1 is likely to be used when raters assess A-level speakers' samples. Therefore, we plan to do a follow-up study with a larger sample: same criteria but 255 speakers representing different proficiency levels, responding to 13 tasks, rated by 20 human raters in spring 2022. Further research is also needed to investigate whether the human raters display bias when using the criteria developed by the research project. Otherwise, the bias of the human ratings may threaten the validity of the automated scoring (see for example Zhang et al. 2020).

In line with our assumption, it proved to be more difficult to receive a certain score on some dimensions, namely Fluency and Range, than others. As McNamara et al. (2019, 75) state, in language assessment contexts, rating scales are usually assumed to have equal steps, although certain scale steps may in fact require more, or less, progress to achieve than others. Moreover, a certain score might reflect a narrower range of abilities than others. In other words, scales do not often advance with equal intervals (McNamara et al. 2019, 75). As shown by this study, a certain score on a certain dimension might require a bigger leap forward in ability than on other dimensions. The results should be taken into account when designing initial report pages that will be shown to the learners. Basically, the goal of this explorative study was to pave way for providing encouraging and accurate automated feedback to learners on their speaking performance.

5. Acknowledgements

We would like to thank the following researchers for their help in collecting human ratings: Ragheb Al-Ghezi, Yaroslav Getman, Ekaterina Voskoboinik from the Aalto University and Heini Kallio from the University of Jyväskylä.

We are grateful for the software engineering students from the University of Helsinki who developed the Moodle plugin for us during spring 2022: Tuomas Alanen, Joona Erkkilä, Topi Harjunpää and Maikki Heijala.

The DigiTala project is funded by the Academy of Finland 2019–2023, and combines expertise in speech and language processing, language education and phonetics at the University of Helsinki (grant number 322619), Aalto University (grant number 322625) and the University of Jyväskylä (grant number 322965).

References

Alanen, T., Erkkilä, J., Harjunpää, T., & Heijala, M. (2022). *Digitala Moodle plugin user manual* (1.0.0). Zenodo. https://doi.org/10.5281/zenodo.6535377

Al-Ghezi, R., Vosboinik, K., Getman, Y., von Zansen, A., Kallio, H., Akiki, C., Kuronen, M., Huhta, A. & Hilden, R. (forthcoming). *Automatic speaking assessment of Spontaneous L2 Finnish and Swedish*.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing* 38(1), 6–40. https://doi.org/10.1177/0265532220927487

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Dordrecht: Springer Science & Business Media.

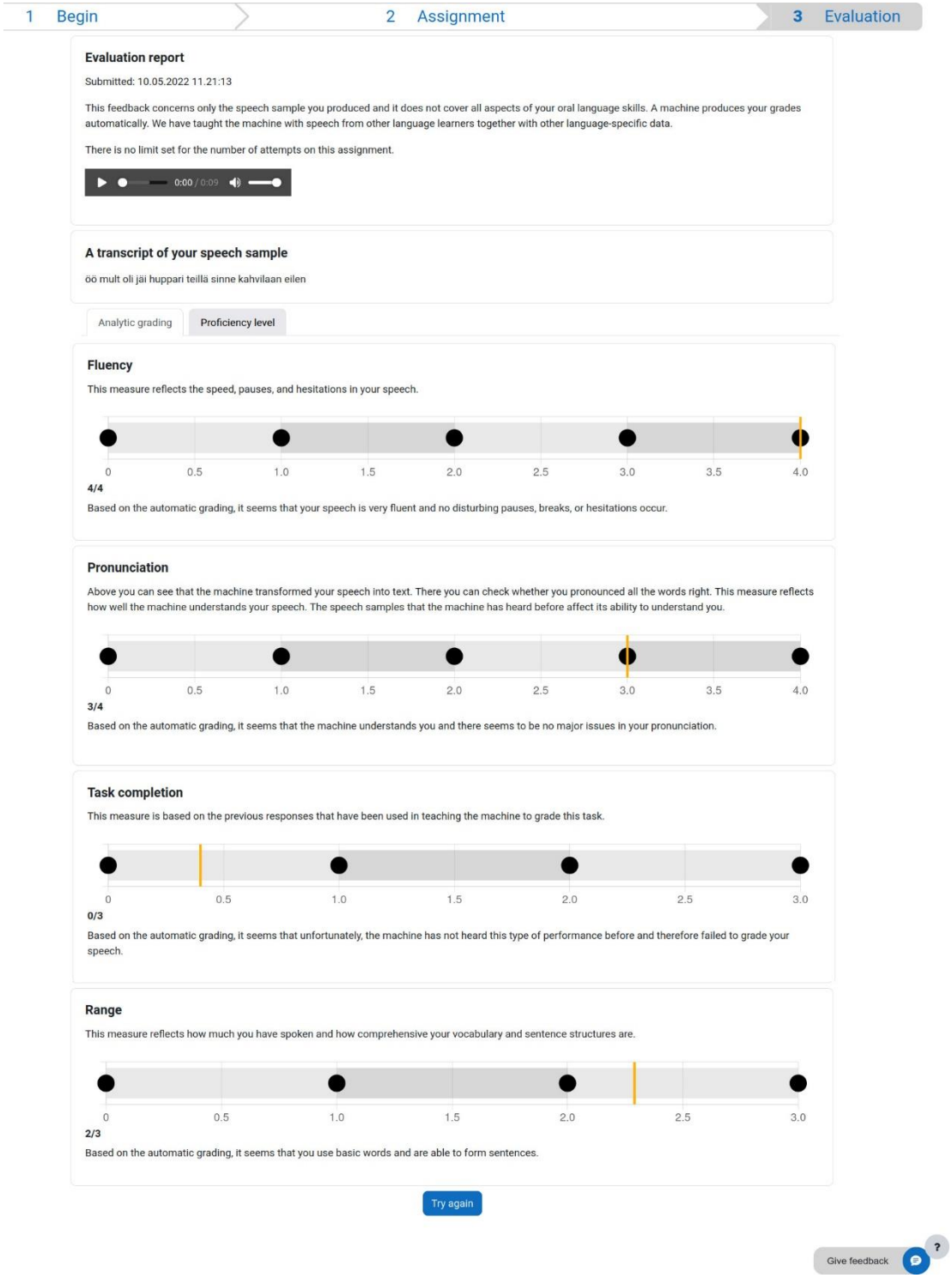
Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education* 162, 104094. https://doi.org/10.1016/j.compedu.2020.104094

Evanini, K., & Zechner, K. (2020). Overview of automated speech scoring. In K. Zechner & K. Evanini (eds.), *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*. New York:

Routledge, 3–20.
Finnish National Agency for Education. (2003). *Lukion opetussuunnitelman perusteet 2003 [National core curriculum for general upper secondary schools 2003]*. https://www.oph.fi/sites/default/files/documents/47345_lukion_opetussuunnitelman_perusteet_2003.pdf
Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer assisted language learning* 27(1), 70–105. https://doi.org/10.1080/09588221.2012.700315
Gu, L., & Davis, L. (2020). Providing SpeechRater Feature Performance as Feedback on Spoken Responses. In K. Zechner & K. Evanini (eds.), *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*. New York: Routledge, 159–175.
Kautonen, M. & von Zansen, A. (2020). DigiTala research project: Automatic speech recognition in assessing L2 speaking. *Kieli, koulutus ja yhteiskunta [Language, Education and Society]* 11 (4). https://www.kieliverkosto.fi/fi/journals/kieli-koulutus-ja-yhteiskunta-kesakuu-2020/digitala-research-project-automatic-speech-recognition-in-assessing-l2-speaking
Linacre J. M. (2002a). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions* 16(2), 878.
Linacre J. M. (2002b). Optimizing rating scale category effectiveness. *Journal of applied measurement* 3(1), 85–106.
Linacre, J. M. (2021). *Facets Rasch measurement* [computer program]. Chicago, IL: Winsteps.com.
McNamara, T., Knoch, U. & Fan, J. (2019). *Fairness, Justice & Language Assessment*. Oxford: Oxford University Press.
Vaarala, H., Riuttanen, S., Kyckling, E., & Karppinen, S. (2021). *Language Reserve. Now! Follow-up on Pyykkö's Report Multilingualism into a strength (2017) : Summary in English*. Jyväskylä: Centre for Applied Language Studies. https://www.jyu.fi/hytk/fi/laitokset/solki/tutkimus/julkaisut/pdf-julkaisut/summary_languagereservenow.pdf
de Vries, B. P., Cucchiarini, C., Bodnar, S., Strik, H., & van Hout, R. (2015). Spoken grammar practice and feedback in an ASR-based CALL system. *Computer Assisted Language Learning* 28(6), 550–576. https://doi.org/10.1080/09588221.2014.889713
Xu, J., Brenchley, M., Jones, E., Pinnington, A., Benjamin, T., Knill, K., Seal-Coon, G. & Geranpayeh, A. (2020). *Linguaskill Building a validity argument for the Speaking test*. Cambridge: Cambridge Assessment English. https://www.cambridgeenglish.org/Images/589637-linguaskill-building-a-validity-argument-for-the-speaking-test.pdf
von Zansen, A., Alanen, T., Al-Ghezi, R., Erkkilä, J., Harjunpää, T., Heijala, M., Kallio, H. (2022). *DigiTala Moodle plugin*. https://github.com/aalto-speech/moodle-puheentunnistus
von Zansen, A. (2022a). *DigiTala's rating criteria: Holistic and analytic scales for assessing L2 speaking. Zenodo*. https://doi.org/10.5281/zenodo.6477089
von Zansen, Anna. (2022b). *DigiTala's speaking tasks for L2 Finnish learners (proficiency level B1)*. Zenodo. https://doi.org/10.5281/zenodo.6562855
von Zansen, Anna. (2022c). *DigiTala's speaking tasks for L2 Finnish learners (proficiency level B2).* Zenodo. https://doi.org/10.5281/zenodo.6562865
von Zansen, A., Kallio, H., Sneck, M., Kuronen, M., Huhta, A., Hilden, R. (accepted). Ihmisarvioijien näkemyksiä suullisen kielitaidon automaattisesta arvioinnista, digitaalisesta arviointiprosessista sekä puhesuorituksista arvioitavista ulottuvuuksista [Human raters' perceptions of the automated assessment of oral language skills, the digital assessment process and the dimensions to be assessed from speaking performances]. In T. Seppälä, S. Lesonen, P. Iikkanen & S. D'hondt (eds.) *AFinLA yearbook*.
von Zansen, A., Sneck, M., & Hilden, R. (accepted a). Lukiolaisten käsitykset ja heidän antamansa palaute suullisen kielitaidon arvioinnista. [Upper secondary school students' perceptions and feedback on automated speaking assessment.] In R. Kantelinen, M. Kautonen, & Z. Elgundi (eds.). *LINGUAPEDA 2021*. Conference Proceedings. Suomen ainedidaktisen tutkimusseuran julkaisuja. Ainedidaktisia tutkimuksia 21.
von Zansen, A., Sneck, M., Hilden, R. (accepted b). *"It was cool and comfortable!"Akateemisten alkeistason S2-opiskelijoiden kokemuksia tietokoneella suoritettavasta puhumisen kokeesta ["It was cool and comfortable!" Academic L2 Finnish learners' perceptions of a computer-based speaking test].* Ainedidaktisia tutkimuksia.
von Zansen, A. & Heijala, M. (forthcoming). *Kielten opettajien ensivaikutelmia suomen ja ruotsin oppijoiden puheen automaattiseen arviointiin kehitetystä työkalusta [Language teachers' first impressions of an automated tool developed for assessing Finnish and Swedish learners' speech]*.
Zhang, M., Bridgeman, B., & Davis, L. (2020). Validity Considerations for Using Automated Scoring in Speaking Assessment. In K. Zechner & K. Evanini (eds.), *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*. New York: Routledge, 21–31.

Appendix 1 Screenshot of the learner's report page



Appendix 2 Results of the MFRM analysis

Table 6.0 All Facet Vertical "Rulers".

Vertical = (1*,2N,3N,4A,S) Yardstick (columns lines low high extreme)= 0,8,-2,4,End

Measr	+Candidate	-Task	-Rater	-Criteria	S.1	S.2	S.3	S.4	S.5
4 +	+	+	+	+	(3)	(4)	(4)	(3)	(4)
	*								
	**								
	*								
	*****								
	*								
3 +	+	+	+	+	+	+	+	+	+
	***								
	***								
	*						---		
	**								
	***					---			
	***								
	***								---
2 +	****	+	+	+	+	+	+	+	+
	*****								
	****								
	****							---	
	***								
	*	13							
	*				---	3	3		3
1 +	***	+	+	+	+	+	+	+	+
	*	20							
			9 13						
	**		3						
	**	3		Fluency Range					
	***		10 12						
		24	14						
	*	16 22	8			---			---
0 *	*	1 6 11 12 19 23 *	*	*	* 2	*	*	2 *	*
	*	2 4 8 10 18 25	5				---		
		5 9 15 17 26	6	Accuracy Task completion					
			2 4						
		14		Pronunciation					
			1 11						
		7	7						
-1 +	+	+	+	+	+	2	+	+	2
	*				---				
							2		

-2 +	+	+	+	+	(1)	(1)	(1)	(1)	(1)
Measr	* = 1	-Task	-Rater	-Criteria	S.1	S.2	S.3	S.4	S.5

S.1: Model = ?,?,?,1,R4 ; Criteria: Task completion
S.2: Model = ?,?,?,2,R4 ; Criteria: Fluency
S.3: Model = ?,?,?,3,R4 ; Criteria: Pronunciation
S.4: Model = ?,?,?,4,R4 ; Criteria: Range
S.5: Model = ?,?,?,5,R4 ; Criteria: Accuracy

Table 7.3.1 Rater Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Rater
1644	582	2.82	2.83	.75	.07	1.05	.8	1.08	1.3	.88	.48	.55	53.5	47.1	13	13
1619	569	2.85	2.84	.73	.07	.90	-1.7	.91	-1.5	1.07	.55	.56	44.0	47.5	9	9
424	149	2.85	2.87	.62	.14	1.10	.8	1.41	3.0	.76	.37	.57	51.8	48.6	3	3
1721	590	2.92	2.97	.40	.07	1.11	1.9	1.06	.9	.96	.48	.54	56.3	50.2	10	10
1638	554	2.96	2.99	.36	.07	1.02	.4	1.06	.9	1.01	.60	.57	61.1	50.5	12	12
1764	591	2.98	3.03	.24	.07	.82	-3.4	.82	-2.9	1.21	.56	.53	61.4	51.0	14	14
1678	547	3.07	3.10	.06	.08	1.04	.7	1.00	.0	1.00	.50	.52	58.5	52.0	8	8
1758	571	3.08	3.15	-.10	.07	.81	-3.3	.92	-1.1	1.12	.57	.53	61.8	52.5	5	5
1867	582	3.21	3.22	-.30	.08	1.14	2.1	1.24	2.8	.88	.54	.51	58.8	53.3	6	6
470	149	3.15	3.21	-.33	.15	.85	-1.2	1.00	.0	1.14	.52	.52	61.9	53.7	4	4
472	149	3.17	3.23	-.38	.15	.84	-1.3	.79	-1.3	1.15	.59	.51	59.8	53.8	2	2
638	201	3.17	3.30	-.63	.13	.82	-1.8	.74	-1.9	1.25	.59	.48	61.4	54.0	1	1
1836	566	3.24	3.31	-.64	.08	.93	-1.0	1.01	.1	1.06	.49	.50	62.0	54.0	11	11
1807	550	3.29	3.35	-.78	.08	1.36	4.8	1.39	3.7	.71	.47	.46	57.7	54.1	7	7
1381.1	453.6	3.05	3.10	.00	.09	.99	-.1	1.03	.3		.52				Mean (Count: 14)	
562.7	185.2	.15	.17	.51	.03	.15	2.2	.20	1.9		.06				S.D. (Population)	
584.0	192.2	.16	.18	.53	.03	.16	2.3	.20	2.0		.06				S.D. (Sample)	

Model, Populn: RMSE .10 Adj (True) S.D. .50 Separation 4.98 Strata 6.97 Reliability (not inter-rater) .96
Model, Sample: RMSE .10 Adj (True) S.D. .52 Separation 5.17 Strata 7.23 Reliability (not inter-rater) .96
Model, Fixed (all same) chi-squared: 492.9 d.f.: 13 significance (probability): .00
Model, Random (normal) chi-squared: 12.6 d.f.: 12 significance (probability): .40
Inter-Rater agreement opportunities: 14606 Exact agreements: 8449 = 57.8% Expected: 7532.1 = 51.6%

Table 7.4.1 Criteria Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N	Criteria
5159	1678	3.07	3.11	.51	.04	.96	-1.4	.95	-1.6	1.07	.62	.59	2	Fluency
2430	999	2.43	2.49	.47	.06	1.01	.1	.99	-.1	.99	.52	.52	4	Range
3354	996	3.37	3.47	-.20	.05	1.08	1.7	1.06	1.0	.92	.53	.57	5	Accuracy
2694	998	2.70	2.79	-.28	.07	1.20	3.3	1.43	4.4	.81	.34	.47	1	Task completion
5699	1679	3.39	3.44	-.51	.04	.91	-2.6	.92	-2.3	1.11	.59	.53	3	Pronunciation
3867.2	1270.0	2.99	3.06	.00	.05	1.03	.2	1.07	.3		.52			Mean (Count: 5)
1321.3	333.5	.38	.38	.42	.01	.10	2.1	.19	2.4		.10			S.D. (Population)
1477.3	372.9	.42	.42	.47	.01	.11	2.4	.21	2.7		.11			S.D. (Sample)

Model, Populn: RMSE .05 Adj (True) S.D. .41 Separation 7.67 Strata 10.56 Reliability .98
Model, Sample: RMSE .05 Adj (True) S.D. .46 Separation 8.59 Strata 11.79 Reliability .99
Model, Fixed (all same) chi-squared: 398.6 d.f.: 4 significance (probability): .00
Model, Random (normal) chi-squared: 4.0 d.f.: 3 significance (probability): .27

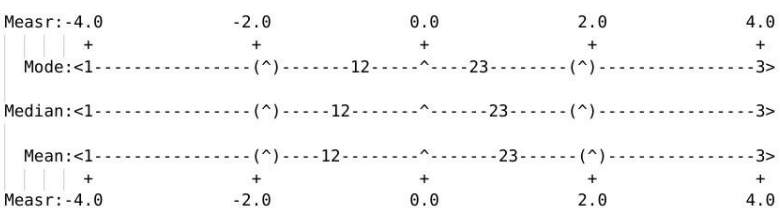
Table 8.1 Category Statistics.

Model = ?,?,?,1,R4 ; Criteria: Task completion

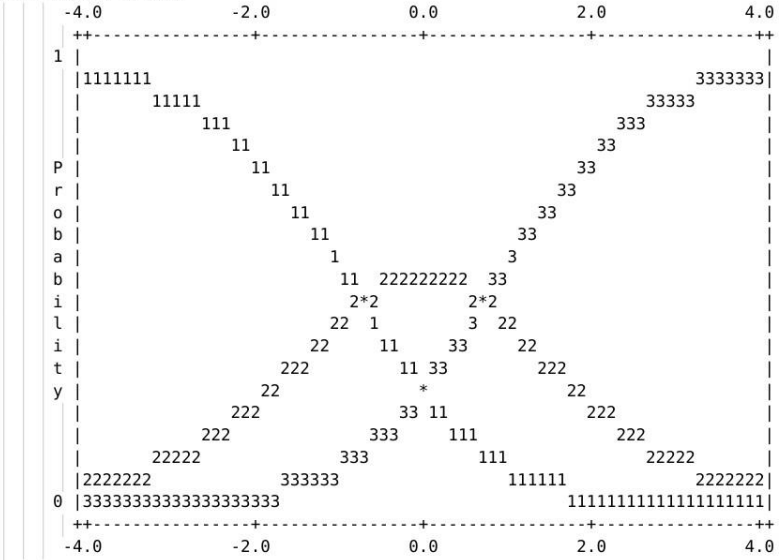
DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts			Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	40	40	4%	4%	.98	.60	1.7			(-1.93)		low	low	100%
2	220	220	22%	26%	1.64	1.42	1.4	-.72	.17	.00	-1.10	-.72	-.90	51%
3	738	738	74%	100%	2.35	2.44	1.1	.72	.08	(1.95)	1.10	.72	.89	100%

(Mean) (Modal) (Median)

Scale structure



Probability Curves



Expected Score Ogive (Model ICC)

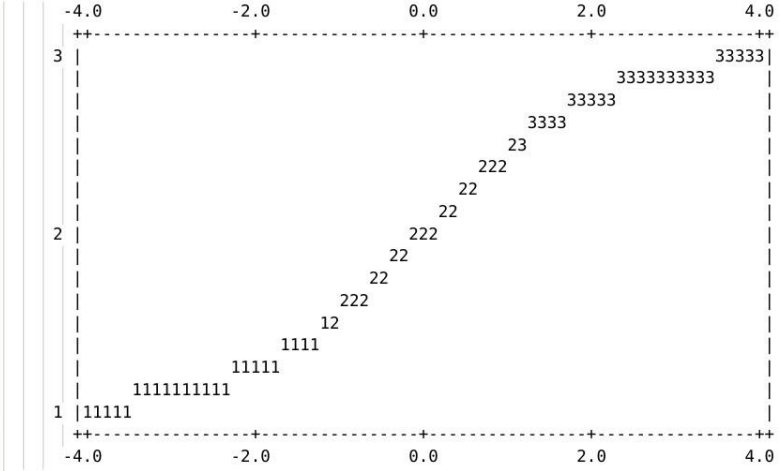


Table 8.2 Category Statistics.

```
 Model = ?,?,?,2,R4  ; Criteria: Fluency
+-------------------------------------------------------------------------------------------------------------+
|           DATA                |   QUALITY CONTROL |RASCH-ANDRICH|  EXPECTATION  |  MOST  |  RASCH-  | Cat|
|      Category Counts       Cum.|  Avge  Exp. OUTFIT| Thresholds  |  Measure at   |PROBABLE| THURSTONE|PEAK|
|Score Total      Used    %    % |  Meas  Meas  MnSq |Measure  S.E.|Category  -0.5 |  from  |Thresholds|Prob|
|-------------------------------+-------------------+-------------+---------------+--------+----------+----|
|  1      35        35    2%   2%|  -.46   -.27   .9 |             |( -3.38)       |   low  |   low    |100%|
|  2     366       366   22%  24%|   .50    .48  1.0 | -2.25    .18|  -1.04   -2.43|  -2.25 |  -2.33   | 63%|
|  3     716       716   43%  67%|  1.29   1.35   .9 |   .24    .07|   1.16     .11|    .24 |    .16   | 55%|
|  4     561       561   33% 100%|  2.25   2.19   .9 |  2.02    .06|(  3.20)   2.33|   2.02 |   2.15   |100%|
+-------------------------------------------------------------------(Mean)---------(Modal)--(Median)------+
```

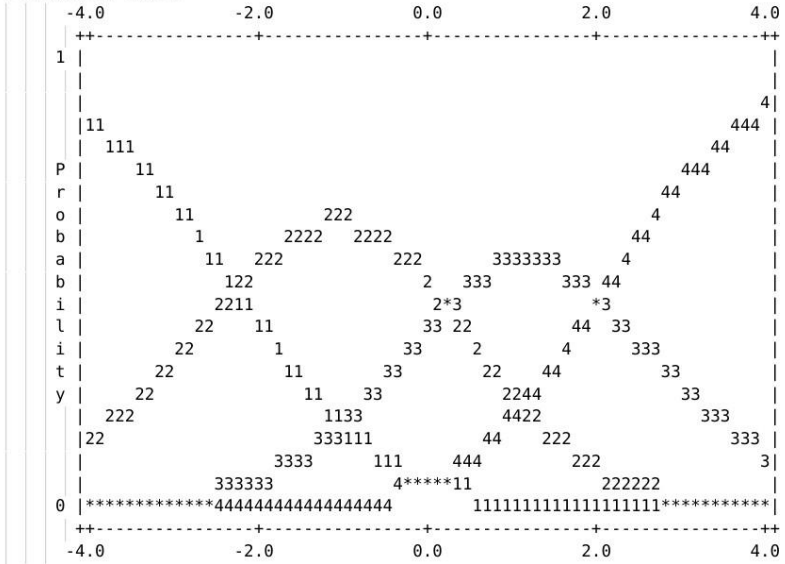
Scale structure

```
Measr:-4.0             -2.0             0.0              2.0              4.0
    +               +               +               +               +
  Mode:<1---(^)-------12---------^---------23-------^-----34--------(^)-----4>

Median:<1---(^)------12----------^--------23--------^------34-------(^)-----4>

  Mean:<1---(^)-----12-----------^--------23--------^--------34-----(^)-----4>
    +               +               +               +               +
Measr:-4.0             -2.0             0.0              2.0              4.0
```

Probability Curves



Expected Score Ogive (Model ICC)

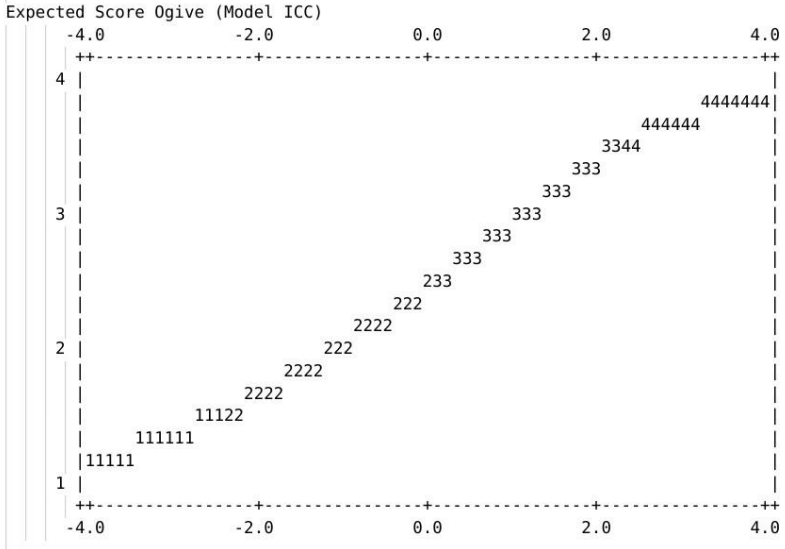


Table 8.3 Category Statistics.

Model = ?,?,?,3,R4 ; Criteria: Pronunciation

DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts			Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	6	6	0%	0%	1.28	.45	1.5			(-3.45)		low	low	100%
2	131	131	8%	8%	1.05*	1.15	.9	-2.30	.42	-1.24	-2.53	-2.30	-2.39	59%
3	737	737	44%	52%	1.97	2.05	.9	-.14	.10	1.17	-.07	-.14	-.11	64%
4	805	805	48%	100%	3.08	3.00	.9	2.44	.06	(3.57)	2.61	2.44	2.50	100%

(Mean) (Modal) (Median)

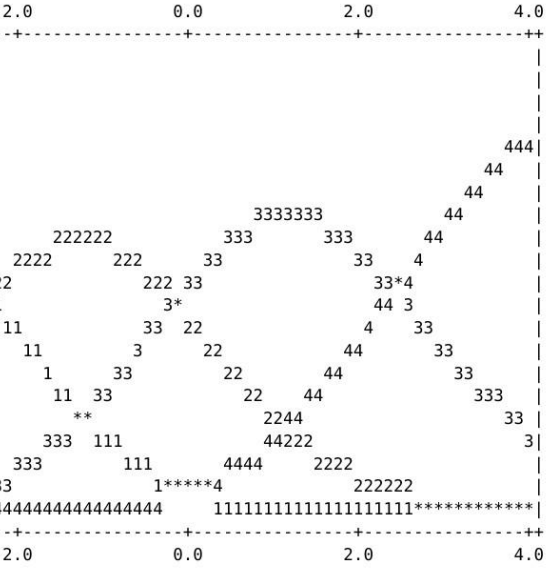
Scale structure

```
Measr:-4.0             -2.0             0.0             2.0             4.0
       +               +                +               +               +
  Mode:<1---(^)------12--------^--------23----------^---------34-------(^)--4>

Median:<1---(^)------12--------^--------23----------^---------34-------(^)--4>

  Mean:<1---(^)-----12---------^--------23----------^----------34------(^)--4>
       +               +                +               +               +
Measr:-4.0             -2.0             0.0             2.0             4.0
```

Probability Curves



Expected Score Ogive (Model ICC)

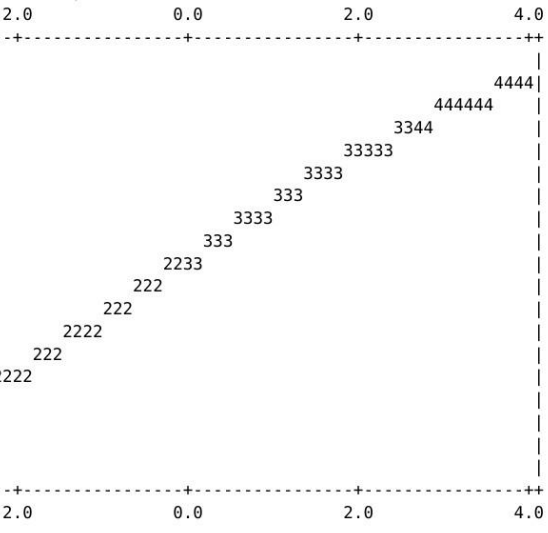
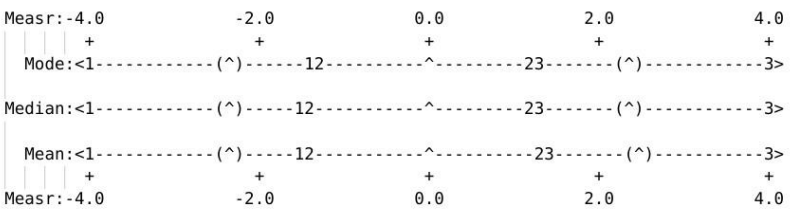


Table 8.4 Category Statistics.

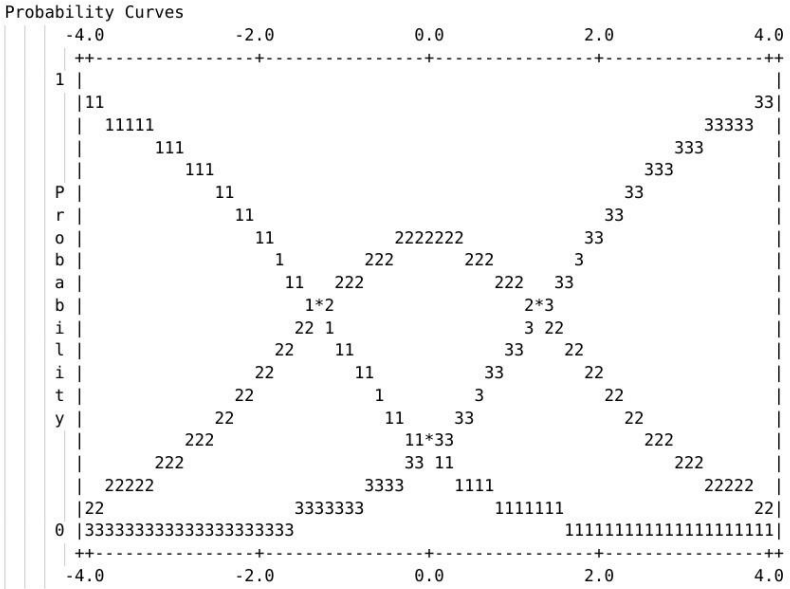
Model = ?,?,?,4,R4 ; Criteria: Range

DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts			Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	70	70	7%	7%	.00	.09	.9			(-2.41)		low	low	100%
2	427	427	43%	50%	1.01	.97	1.0	-1.29	.13	.00	-1.45	-1.29	-1.36	65%
3	502	502	50%	100%	1.91	1.93	1.0	1.29	.07	(2.42)	1.46	1.29	1.35	100%
(Mean)										(Modal)		(Median)		

Scale structure



Probability Curves



Expected Score Ogive (Model ICC)

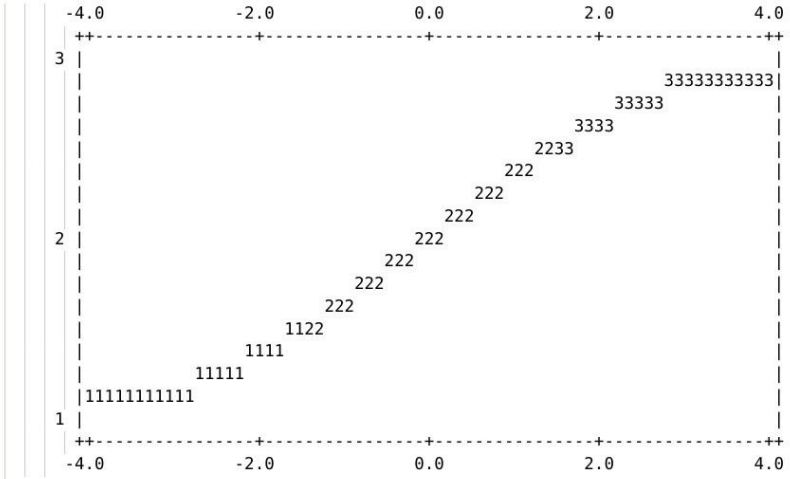


Table 8.5 Category Statistics.

Model = ?,?,?,5,R4 ; Criteria: Accuracy

DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category Counts			Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	9	9	1%	1%	.46	.28	1.2			(-3.22)		low	low	100%
2	130	130	13%	14%	.99	.90	1.1	-2.09	.34	-.94	-2.28	-2.09	-2.17	62%
3	343	343	34%	48%	1.72	1.72	.9	.32	.10	1.09	.14	.32	.22	50%
4	514	514	52%	100%	2.59	2.61	1.1	1.77	.07	(3.00)	2.16	1.77	1.94	100%
									(Mean)		(Modal)	(Median)		

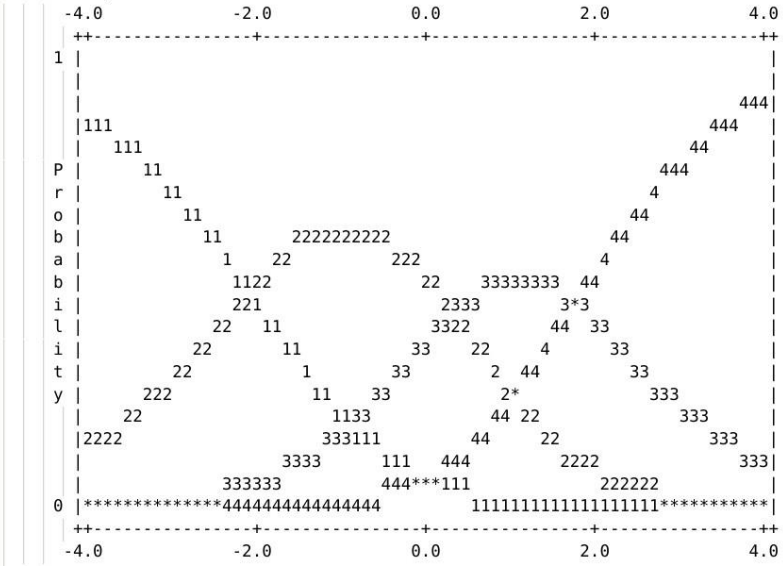
Scale structure

```
Measr:-4.0             -2.0             0.0              2.0              4.0
       +                +                +                +                +
  Mode:<1-----(^)------12---------^---------23-----^----34--------(^)-------4>

Median:<1-----(^)------12---------^--------23------^------34------(^)-------4>

  Mean:<1-----(^)-----12----------^-------23-------^-------34-----(^)-------4>
       +                +                +                +                +
Measr:-4.0             -2.0             0.0              2.0              4.0
```

Probability Curves



Expected Score Ogive (Model ICC)

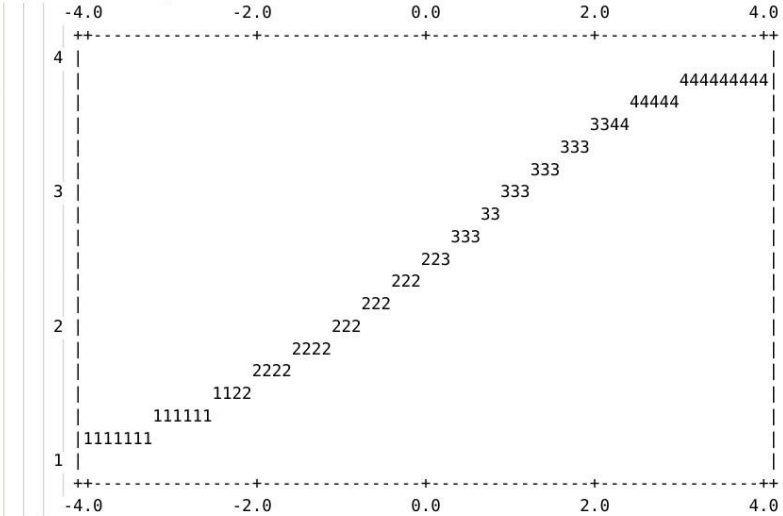


Table 4.1 Unexpected Responses (61 residuals sorted by u).

Cat	Score	Exp.	Resd	StRes	Nu	Ca	Nu	Ta	Nu	Ra	N	Criteria	Sequence
2	2	3.0	-1.0	-7.9	36	36	25	8c	11	11	1	Task completion	6057
2	2	3.0	-1.0	-6.4	43	43	2	2	7	7	1	Task completion	1180
2	2	3.0	-1.0	-6.2	68	68	2	2	7	7	1	Task completion	1195
1	1	2.9	-1.9	-5.7	6	6	2	2	4	4	1	Task completion	745
1	1	2.9	-1.9	-5.6	37	37	26	8d	6	6	1	Task completion	6177
2	2	3.0	-1.0	-5.3	69	69	25	8c	6	6	1	Task completion	5932
1	1	2.9	-1.9	-5.1	39	39	11	3b	11	11	1	Task completion	4056
2	2	3.0	-1.0	-5.0	49	49	26	8d	6	6	1	Task completion	6182
1	1	2.8	-1.8	-4.9	24	24	26	8d	12	12	1	Task completion	6277
1	1	2.8	-1.8	-4.8	18	18	2	2	3	3	1	Task completion	700
2	2	3.0	-1.0	-4.6	12	12	23	8a	5	5	1	Task completion	5458
3	3	3.9	-.9	-4.2	36	36	9	1f	11	11	3	Pronunciation	3689
1	1	2.8	-1.8	-4.2	59	59	10	3a	7	7	1	Task completion	3796
2	2	2.9	-.9	-4.1	16	16	2	2	6	6	1	Task completion	995
2	2	3.8	-1.8	-4.1	43	43	4	6b	12	12	5	Accuracy	2779
1	1	2.8	-1.8	-4.0	3	3	10	3a	6	6	1	Task completion	3761
2	2	3.8	-1.8	-4.0	68	68	4	6b	12	12	5	Accuracy	2784
1	1	3.3	-2.3	-3.9	10	10	2	2	7	7	3	Pronunciation	1097
2	2	2.9	-.9	-3.9	19	19	25	8c	11	11	1	Task completion	6042
1	1	3.3	-2.3	-3.9	29	29	7	1c	13	13	3	Pronunciation	3384
1	1	2.8	-1.8	-3.9	34	34	11	3b	5	5	1	Task completion	3949
1	1	2.8	-1.8	-3.8	17	17	2	2	3	3	1	Task completion	695
2	2	3.8	-1.8	-3.8	68	68	23	8a	12	12	5	Accuracy	5642
2	2	3.7	-1.7	-3.7	5	5	15	3f	7	7	3	Pronunciation	4374
2	2	3.7	-1.7	-3.7	12	12	2	2	3	3	3	Pronunciation	677
1	1	2.7	-1.7	-3.7	14	14	12	3c	1	1	1	Task completion	4107
1	1	2.7	-1.7	-3.7	25	25	2	2	3	3	1	Task completion	735
2	2	3.7	-1.7	-3.7	52	52	24	8b	7	7	3	Pronunciation	5760
2	2	2.9	-.9	-3.7	54	54	2	2	7	7	1	Task completion	1190
2	2	2.9	-.9	-3.6	18	18	2	2	6	6	1	Task completion	1005
2	2	3.7	-1.7	-3.5	1	1	2	2	1	1	5	Accuracy	474
2	2	3.7	-1.7	-3.5	23	23	4	6b	7	7	3	Pronunciation	2527
1	1	3.1	-2.1	-3.5	39	39	11	3b	13	13	3	Pronunciation	4074
1	1	2.7	-1.7	-3.5	53	53	16	4a	14	14	4	Range	4527
2	2	3.7	-1.7	-3.4	1	1	6	1b	7	7	3	Pronunciation	3108
2	2	2.9	-.9	-3.4	6	6	2	2	7	7	1	Task completion	1085
1	1	2.7	-1.7	-3.4	10	10	2	2	7	7	1	Task completion	1095
1	1	3.3	-2.3	-3.4	10	10	2	2	7	7	5	Accuracy	1099
2	2	2.9	-.9	-3.4	12	12	2	2	3	3	1	Task completion	675
2	2	2.9	-.9	-3.4	16	16	19	4d	5	5	1	Task completion	4904
2	2	2.9	-.9	-3.4	31	31	25	8c	10	10	1	Task completion	6037
1	1	2.7	-1.7	-3.4	32	32	3	5	8	8	1	Task completion	2131
2	2	3.7	-1.7	-3.4	41	41	19	4d	6	6	5	Accuracy	4943
1	1	3.0	-2.0	-3.3	29	29	1	1d	13	13	3	Pronunciation	433
1	1	2.6	-1.6	-3.3	29	29	17	4b	7	7	4	Range	4584
1	1	2.7	-1.7	-3.3	64	64	23	8a	6	6	1	Task completion	5493
2	2	2.9	-.9	-3.2	24	24	24	8b	7	7	1	Task completion	5748
2	2	2.9	-.9	-3.2	27	27	2	2	7	7	1	Task completion	1170
2	2	3.7	-1.7	-3.2	32	32	10	3a	6	6	5	Accuracy	3775
2	2	3.7	-1.7	-3.2	55	55	2	2	10	10	5	Accuracy	1554
1	1	2.7	-1.7	-3.2	64	64	25	8c	5	5	1	Task completion	5927
2	2	3.7	-1.7	-3.2	66	66	23	8a	6	6	5	Accuracy	5502
2	2	3.6	-1.6	-3.1	8	8	7	1c	8	8	3	Pronunciation	3290
1	1	3.2	-2.2	-3.1	15	15	10	3a	7	7	5	Accuracy	3785
1	1	2.6	-1.6	-3.1	16	16	2	2	9	9	4	Range	1358
2	2	2.9	-.9	-3.1	18	18	14	3e	12	12	1	Task completion	4298
3	3	3.9	-.9	-3.1	21	21	7	1c	5	5	3	Pronunciation	3250
1	1	2.6	-1.6	-3.0	7	7	4	6b	4	4	1	Task completion	2380
3	3	3.9	-.9	-3.0	36	36	9	1f	11	11	2	Fluency	3688
2	2	2.9	-.9	-3.0	41	41	2	2	6	6	1	Task completion	1060
2	2	2.9	-.9	-3.0	53	53	18	4c	12	12	1	Task completion	4862
Cat	Score	Exp.	Resd	StRes	Nu	Ca	Nu	Ta	Nu	Ra	N	Criteria	Sequence

What Teaching an Algorithm Teaches When Teaching Students How to Write Academic Texts

Michael Pace-Sigge, Dian Toar Sumakul
University of Easter Finland, Finland; Universitas Kristen Satya Wacana, Indonesia
E-mail: michp@uef.fi

Abstract

In April 2019, Springer Heidelberg published the first ever AI-written academic text-book (Beta Writer 2019). The people who developed the algorithm thus became instructors of "how to write an academic text". This can be seen to mirror the task of instructors in EAP writing classes. This paper sets out to investigate whether the Springer editors were successful in turning their plan into something that compares well to natural occurring text (Hyland and Tse, 2007; Pace-Sigge 2018). In this article, we employ the Beta Writer book as a computer-simulation of what 'teaching academic writing' is like.
Using a corpus-based analysis, the Beta Writer book is directly compared to 10 other textbooks covering the same topics. Words and phrases typical of academic writing were directly compared for their occurrence, grammatical structure and prosodies. Our research shows that the AI algorithm developed for the publisher, has created a text comprehensible to a human reader: this, in turn, indicates what factors would be useful for a human teacher to take into consideration when teaching human students. However, it is also the case that often, the result is non-natural and unsuitable in Beta Writer: this highlights areas that need particular attention when teaching academic writing.

Keywords: Artificial Intelligence (AI), academic writing, corpus linguistics, EAP, lexical priming, natural language

1. Introduction

1.1. What this Paper sets out to do

Springer Heidelberg published the first ever AI-written academic text-book (Beta Writer 2019) where those who developed the algorithm would have used a preconception of what an academic text looks like and, in this, they became instructors; teaching the algorithm, they created "how to write an academic text". This, to us, can be seen as mirroring the task instructors in EAP writing classes have.

This paper sets out to investigate whether the Springer editors were successful in turning their plan into something that compares well to naturally occurring text (cf. Hoey 2005). We aim to highlight in which areas the algorithm has created a text that will be acceptable to a human reader: this, in turn, would indicate moves that would be useful to undertake when teaching in a classroom. However, areas where the AI-created text falls short of meeting expectations allow us to highlight points that need particular attention when teaching academic writing.

1.2. The Role of AI in Teaching

Early AI tools like the Intelligent Tutoring System (ITS) were introduced into language classrooms in the 1980s. The goal of the early ITS was to simulate the teaching and learning processes between a student and a teacher (Seidel & Park 1994). Self (1998) explains that the early systems were focused on adapting to the needs of the learners and the technology has been found to be particularly beneficial to students learning a language. In its early versions, AI was reported to be able to check students' grammar and provide sophisticated feedback (Bailin 1987), engage learners in a written dialogue (Jehle 1987), process students' language input (Holland et al. 1993), and provide more effective feedback in grammar lessons (Nagata 1996).

More recently, AI is used to analyse, comprehend, and produce human language (Lu 2018).

These abilities are integrated into different applications (apps) to help students learn. For example, AI has been found to be able to improve students' academic achievements (Köse & Arslan 2015), provide meaningful communications (Lu 2018), improve students' confidence in learning a foreign language (Haristiani 2019), increase students' reading comprehension (Bailey *et al*. 2021), enhance listening skills (Ghoneim & Elghotmy 2021), improve speaking performance (El Shazly 2021), and assist students in their writing and help build students' creativity (Sumakul 2019; Sumakul et al. 2021).

Nevertheless, there have also been critics of this approach from early on. The promises offered by AI for language teaching and learning were judged to be misunderstood (Last 1989), not efficient (Wolff 1993), and exaggerated (O'Brien 1993). There have been doubts that AI can bring advantages to language classrooms (Salaberry 1996); or, if advantages are to be found, these would only have a moderate impact on students' attainment (Steenbergen-Hu & Cooper 2014). Indeed, a recent study by Gallacher et al. (2018) claims that AI is not a legitimate language learning tool.

Consequently, the question of whether to use AI in a classroom must be reframed. It is not solely about the technology chosen but rather, the didactic approach that is being applied. Yet this weak connection to theoretical pedagogical perspectives has been found as one of the crucial issues preventing the widespread use of AI in education by Rieland (2017) and Zawacki-Richter (2019).

1.3. The Example of Training AI to Write a Textbook

AI - Artificial Intelligence - can have a wide range of meanings. It should therefore not come as a surprise that calling *BetaWriter* "the first academic textbook written by AI" can perhaps be misunderstood[80]. The editors of Springer's book stress several times that they "have chosen a conservative approach". This means that the algorithm developed to create the textbook was fed 1,000 recent research articles which all deal with lithium-ion batteries. The approach chosen for this book highlights the very strong lead taken by the editors: the design was directed to focus on a given set of key-words – rather than using a machine-learning approach where it is the algorithm that trawls a wide range of texts and mines it for relevant data. Data-mining, given sufficient resources, can uncover unexpected material and connections (cf. Manning 2015). Crucially, the processing of the source material has had a complete focus on the textual element. This means that anything beyond that - tables, graphs and images - was both ignored and subsequently did not appear in the finished product. This is important, as comparable books have a high content of non-text elements and these are clearly important to the comprehension of the context. This can be seen in the comparison of a random page in Figure 1.

Collins (2019) refers to *Beta Writer* as a "low-level literature review by doing no more than summarising the contents of existing papers" several other reviewers make this point as well, and a number or reviewers refer to Jeff Bigham (in Claburn 2019): "the very nature of extractive summary means it will be coherent in chunks, so long as the input texts are coherent".

The question arises, therefore, of whether the form of a book is appropriate and suitable. Springer's PR says that *Beta Writer's*[81] book (in the following, BW) provides a response to information overload. Being simply a searchable e-text as Day (2019) points out that "the algorithm's most likely use-case (…) is to trawl through literature on a given topic".

[80] A summary of the architecture to construct this book is kept brief here: all details can be found in Chapter 1.2 of *Lithium Ion Batteries* which is free to download.
[81] This is often referenced as "Writer, B." in references. However, this misses the idea behind the name. Therefore, this article will refer to the name given by Springer editors, viz. *Beta Writer*. It is a moot point whether this is meant to be a pun – *beta* version can be pre-release version and *beta-betta= better* writer.

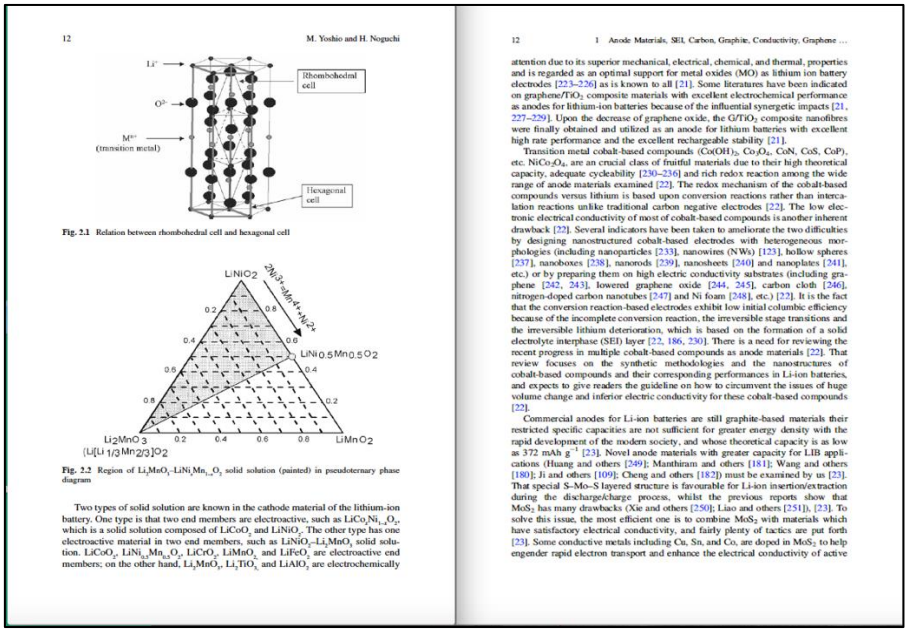


Figure 1: Page 12 of *Lithium-Ion Batteries* (Yoshio, Brodd & Kozawa, eds., 2009): left; *Lithium-Ion Batteries* (Beta Writer 2019): right.

2. Analysis of the Beta Writer Text

2.1. Research Methodology

In order to see in how far the AI-written text reflects natural language in terms of its lexical choices, grammar and word connotations, the book is contrasted with prototypical usage as represented by a small corpus of eleven textbooks in the field, published between 2009 and 2018[82]. As mentioned, this corpus is the *Lithium-Ion_Batteries_Books* (LIBB in the following). Initially, a keyword analysis was undertaken, using WordSmith 8 (Scott 2020) to observe those items which are over-used in the *Beta Writer* text as well as those which are overused in LIBB. The next step was to select those words for the investigation which are not subject-typical but seen as typical of academic writing in general, based on the definitions given by Carter and McCarthy (2006). The final word-choice was then checked for statistical significance, employing Rayson's (2016) log-likelihood calculator. Finally, the items of particular interest were investigated in their occurrence in multi-word clusters.

2.2. Single-item Analysis

2.2.1. Connectors

Looking at the most frequent words used as connectors, it stands out that *and, which because* and *though* are strongly and significantly over-used; the opposite is true for *but, or, also, however, although* and *upon* as Table 1 shows. It must be noted that Carter and McCarthy (2006) refer to these last three words as "typical of academic writing".

[82] Full details of the LIBB corpus can be found here: https://sites.uef.fi/pathwaystotextualitysymposium/external-materials_pace-sigge/

	observed frequencies		expected frequencies		
connectors	LIBB	BW	LIBB	BW	Log-likelihood
THOUGH	81	87	155.23	12.77	228.51
WHICH	3620	556	3858.60	317.40	161.27
AND	34161	3252	34569.43	2843.57	60.89
BECAUSE	1016	127	1056.13	86.87	17.74
ALSO	2529	42	2375.59	195.41	187.37
HOWEVER	1671	16	1558.78	128.22	165.74
OR	4008	146	3838.28	315.72	121.64
BUT	1508	23	1414.64	116.36	118.18
ALTHOUGH	387	3	360.36	29.64	41.46
UPON	406	9	383.46	31.54	23.81
Corpus size	1,324,535	108,952			

Table 1: Contrastive use of *connectors*

What stands out is that *though* is significantly over-used, while *although* is hardly in evidence in BW.

	observed frequencies		expected frequencies		
adjuncts	LIBB	BW	LIBB	BW	Log-likelihood
THUS	875	0	808.50	66.50	138.33
THEREFORE	833	0	769.69	63.31	131.69
THEN	727	14	684.68	56.32	48.23
Corpus size	1,324,535	108,952			

Table 2: Contrastive use of *resultative adjuncts*

The adjuncts in Table 2 are typically used to describe causation (Carter & McCarthy 2006). It can be seen that the most frequent forms in LIBB are significantly underused in BW[83].

2.2.2. Boosters and hedges

Looking at adverbs and adjectives, a number of highly key items appear in BW frequently yet are comparatively less used in human-written textbooks, namely, *high, low, greater, maximal, fruitful, systemic,* as well as *efficiently, substantially, satisfactory*. It must be noted that BW appears to have a preference for words that boost the head-word. This is all the more interesting as Figure 4 shows that words typically seen as boosters in LIBB are underused in BW.

As can be seen in Figure 4, neither set uses boosters particularly frequently. The level of statistically relevant divergence is lower than for the words discussed above. What stands out is that this set of highly conventionalised boosters are fairly typical in LIBB, none more so than *very* (LL=128). Table 3, by contrast, shows the type of boosters which cannot be described as conventionalised in academic writing.

[83] In the whole book, *thus* appears 8 times and *therefore* 4 times: these are only found in the human-written foreword.

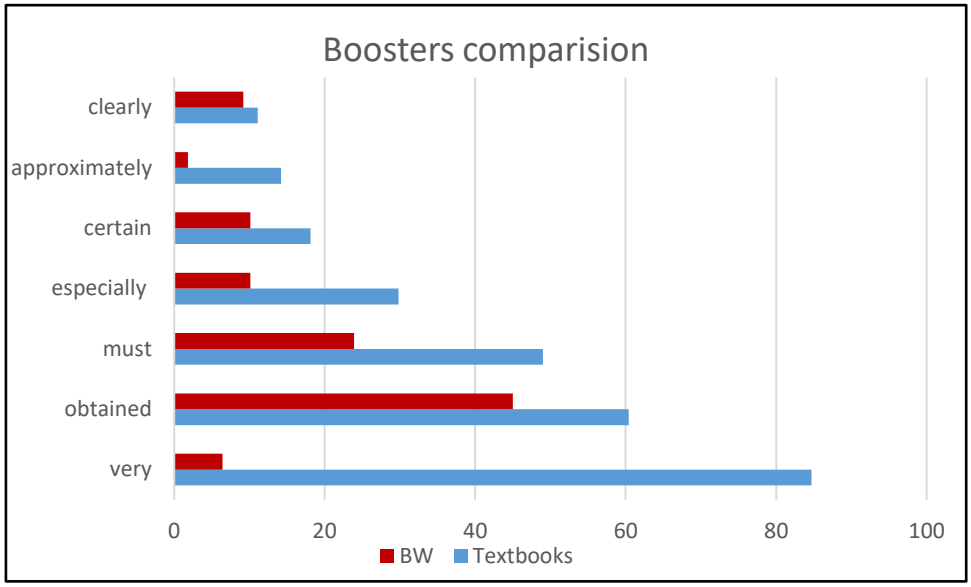


Figure 4: Boosters comparison: LIBB high

	observed frequencies		expected frequencies		
	LIBB	BW	LIBB	BW	Log-likelihood
GREATER	159	126	263.34	21.66	283.27
SATISFACTORY	33	116	137.68	11.32	445.50
FRUITFUL	4	72	70.22	5.78	340.37
EFFICIENTLY	19	71	83.16	6.84	276.16
SYSTEMIC	4	53	52.67	4.33	244.82
SUBSTANTIALLY	48	49	89.63	7.37	125.67
Corpus size	1,324,535	108,952			

Table 3: *Boosters* comparison: BW high

As can be seen, most of these adverbs and adjectives are very rare in the textbooks corpus (LIBB) – this seems to indicate that the AI-writer tried to replace highly suitable words with near-synonyms that might not be fully appropriate for this text-type.

	observed frequencies		expected frequencies		
	LIBB	BW*	LIBB	BW	log-likelihood
MAY	1002	3	928.62	76.38	133.00
CAN	3745	187	3633.15	298.85	51.77
WILL	1417	43	1349.03	110.97	57.77
MUST	649	23	620.92	51.08	20.70
SHOULD	551	8	516.51	42.49	44.51
COULD	692	142	770.61	63.39	80.14
MIGHT	139	55	179.26	14.74	74.10
Corpus size	1,324,535	108,952			

Table 4: *Modals acting as hedges* comparison (N excludes non-AI constructed text)

The most interesting divergence demonstrated by Table 4 is clearly that *may* and *can* are underused, yet related (epistemic) modals – *could* and *might* – appear in the statistical calculation as significantly overused. This, again, needs to be investigated further. As Carter and McCarthy (2006, 279) highlight "since academic discourse is often about theories (…) hedging is very important in academic styles. Less often, it is also necessary to assert a claim or viewpoint quite directly (…), a process we shall

refer to as boosting". This points towards strong personal input by the respective writer(s) who need to assess whether claims made are tentative on the one hand or can be given with more force and confidence on the other.

2.3. Item Behaviour in Longer Clusters

2.3.1. The case of strong boosterism

Starting with the ten lines of *clearly* in BW[84], we find *is clearly* being more typical in LIBB, yet, crucially, BW speaks of *clearly suggested / confirmed / decided* / *indicate*: yet none of these bigrams can be found in LIBB. By contrast, the most frequent LIBB fixed form, *clearly observed* (1/10 of *clearly* uses) or *clearly identified (4/10)* does not occur in BW either. Thus, it appears that *clearly* tends to be used to state facts in LIBB while BW employs it with some degree of uncertainty. In short, collocations and, to a degree, semantic prosodies appear to diverge.

Must occurs 49 times and *must be* is the most frequent bigram in both sources. Beyond that the structure mirrors what we have seen with *clearly* – the pre-modified adjectives do not correspond at all, meaning that even firmly fixed phrases like "must also (be)" which appear 1/30 times in LIBB are not found in BW. Instead, like with *clearly*, *must* can be found post-modified with *quite.* Looking at prosodies, "must have" indicates absolutes in BW such as "must have superior" and "must have 95.5% purity". In LIBB, by contrast, the less extreme *must have (a) good…* is used as well as *must have a low / a basic / the same* used.

This structure becomes even more apparent in the 51 concordance lines of *substantially* in BW: 1/5 occurrences are *substantially enhanced*, with over 1/9 being *can substantially enhance.* The colligational structure in LIBB is different here. The hedge appears only once – "can vary substantially" yet the word itself is often followed by a comparative (larger, longer, higher, lower, safer, more).

The only booster from Table 3 that also appears with a relatively high number in both LIBB and BW is the comparative *greater*. Proportionally that is 12 per 100,000 words compared to 115,6 per 100,000 words. The collocates for *greater* are often the same with similar frequencies. Both sources use the phrase *(is) greater than that of* – 1/40 in LIBB, yet a far more frequent 1/14 in BW. By contrast, the use of *greater extent* is found only once in BW, yet four times in LIBB. More importantly, the mathematical formulation *the greater the* appears nine times in LIBB – but is not to be found in BW. Different, too, are the pre-modifiers. Amongst the most frequent in LIBB are *much, even, slightly, significantly* and *substantially* as the most frequent. Beta Writer also uses connectors frequently – so we have *and (a) greater* 16 times (in LIBB: 3) *with greater* appears ten times (3). There is also a preference for words referring to observations in BW: *display(s)* (6), *showed* (4), *indicate* (2), whereas LIBB records none of these only uses *is.*

2.3.2. Strong divergence for hedges and causation markers

Table 4 highlights that there is a clear discrepancy between the modals employed between the two sources, with most underused in BW, yet *could* and *might* significantly overused.

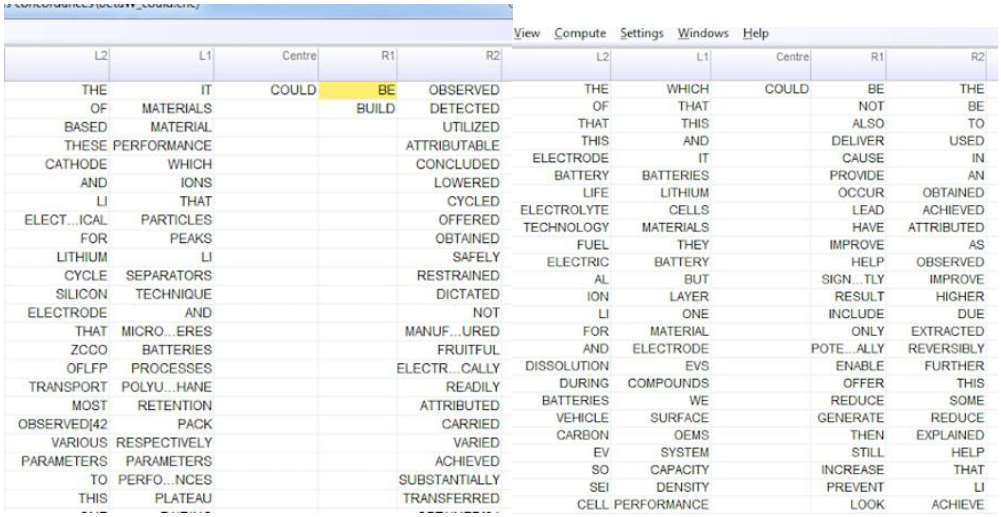
There are marked difference in usage found here. *May* occurs 1002 times in LIBB. In BW, *may* occurs 13 times – 10 of which are in the human-written introduction. One could assume therefore, that *might* covers both uses found in LIBB. However, that is not the case. In fact, the clear

[84] *Certain* and *especially* are equally infrequent in BW (11 occ.).

characteristic of *might* in BW seems to have a preference for combinations into long clusters. So, for example, this thrice –occurring tapeworm of a sentence "the electrochemical properties of ZCO might be linked to the electrochemical performance of ZCCO" which appears in three different parts of the book and is nested between wordings with little variation. There is another of these tapeworms occurring twice and at least one more saying the same using a slightly different word order.

There is just one clear point of overlap, namely the trigram *which might be* which occurs 1/7 in BW and 1/13 in LIBB, followed by the trigram *it might be* which appears with a similar proportional frequency. *Might not be* is not recorded in BW, yet occurs six times in LIBB. In LIBB, furthermore, there is strong indication for potential causation - *might cause* occurs nine times, and might *lead/result* three times. Each of these occurs only once in BW. The contrast BW provides using *might* as a weak hedge with a less direct marker of causation: *might be linked to; might give rise to.* There is one further crucial difference: BW leans to negative prosody, with the L1 word *uncertainties* and the R1 words *not, suffer, worsen, severely, downturn, decline* while there is a lack of positive terms. LIBB also has negative terminology in R1: *impair, kill, decrease, damage, destroy, impede.*

The modal *could* appears significantly overused in BW compared to the textbooks. It is also hyper-formulaic. In the 142 concordance lines, only one is *could build* whereas all others are *could be,* which is also the most prominent form in LIBB – yet here it appears in under 4/10 concordance lines with "could not" being second-most frequent at 6/100 lines. Further variations are shown in Figure 6.



L2	L1	Centre	R1	R2
THE	IT	COULD	BE	OBSERVED
OF	MATERIALS		BUILD	DETECTED
BASED	MATERIAL			UTILIZED
THESE	PERFORMANCE			ATTRIBUTABLE
CATHODE	WHICH			CONCLUDED
AND	IONS			LOWERED
LI	THAT			CYCLED
ELECT...ICAL	PARTICLES			OFFERED
FOR	PEAKS			OBTAINED
LITHIUM	LI			SAFELY
CYCLE	SEPARATORS			RESTRAINED
SILICON	TECHNIQUE			DICTATED
ELECTRODE	AND			NOT
THAT	MICRO...ERES			MANUF...URED
ZCCO	BATTERIES			FRUITFUL
OFLFP	PROCESSES			ELECTR...CALLY
TRANSPORT	POLYU...HANE			READILY
MOST	RETENTION			ATTRIBUTED
OBSERVED[42	PACK			CARRIED
VARIOUS	RESPECTIVELY			VARIED
PARAMETERS	PARAMETERS			ACHIEVED
TO	PERFO...NCES			SUBSTANTIALLY
THIS	PLATEAU			TRANSFERRED

View Compute Settings Windows Help

L2	L1	Centre	R1	R2
THE	WHICH	COULD	BE	THE
OF	THAT		NOT	BE
THAT	THIS		ALSO	TO
THIS	AND		DELIVER	USED
ELECTRODE	IT		CAUSE	IN
BATTERY	BATTERIES		PROVIDE	AN
LIFE	LITHIUM		OCCUR	OBTAINED
ELECTROLYTE	CELLS		LEAD	ACHIEVED
TECHNOLOGY	MATERIALS		HAVE	ATTRIBUTED
FUEL	THEY		IMPROVE	AS
ELECTRIC	BATTERY		HELP	OBSERVED
AL	BUT		SIGN...TLY	IMPROVE
ION	LAYER		RESULT	HIGHER
LI	ONE		INCLUDE	DUE
FOR	MATERIAL		ONLY	EXTRACTED
AND	ELECTRODE		POTE...ALLY	REVERSIBLY
DISSOLUTION	EVS		ENABLE	FURTHER
DURING	COMPOUNDS		OFFER	THIS
BATTERIES	WE		REDUCE	SOME
VEHICLE	SURFACE		GENERATE	REDUCE
CARBON	OEMS		THEN	EXPLAINED
EV	SYSTEM		STILL	HELP
SO	CAPACITY		INCREASE	THAT
SEI	DENSITY		PREVENT	LI
CELL	PERFORMANCE		LOOK	ACHIEVE

Figure 6: L2, L1 and R1 patterns for BW (left) and LIBB (right)

This makes *could* a prime example of divergence. The colligation pattern shows no other use but the be-infinitive[85]. No other verbs are employed, nor are there connectors (*also*), time-markers (*then*), adverbs (*only*), prepositions (*in*), relative-clause markers (*that*) which are all found in LIBB. Crucially, the colligational and prosodical impact of having 6 per cent of all uses of *might* negated as in the textbooks marks a significant divergence[86].

Looking at the verbs the modal pre-modifies in R2-position, there is a significant preference for words referring to observations in BW: *could be detected / observed / utilized / concluded / offered.* While *could be observed* occurs in LIBB, too, the last two are not recorded in LIBB. By contrast,

[85] This is true as *could build* only appears a single time.
[86] LL of 5.2 when the whole of the corpus is taken into account; LL of 15.5 for the use of *could* only.

could be due to is not found in BW and it must be noted that there is a shift in word class, too; 'attributed' becomes 'attributable' (V to ADJ).

Another structural shift is shown in Figure 6: in LIBB, the typical construction appears to be a relative clause: *which could* and *that could* appear in almost 7/100 cases, yet the former occurs only half as often and *that could* even less frequently in BW[87]. All these points together highlight that *could*, overall, appears to carry different lexical functions in BW.

Can be is the prototypical structure in LIBB – found in almost 54/100 concordance lines but only in 12/100 in the AI-written part of the BW. The textbooks employ *can be used* in almost 5/100 lines – this cluster is absent in BW. Instead, one finds "can (substantially) enhance" which are very rare in LIBB. Like *could*, *can* is typically part of a relative clause.

While LIBB in R2 shows use of verbs like *expressed, considered, described* and *observed,* these are not in evidence in BW. Crucially, *provide, cause, result* and *use* are used in LIBB. This colligational link to causality does not appear to exist in BW.

When it comes to causality writers can use a variety of terms to describe cause-effect situations, see Figure 7.

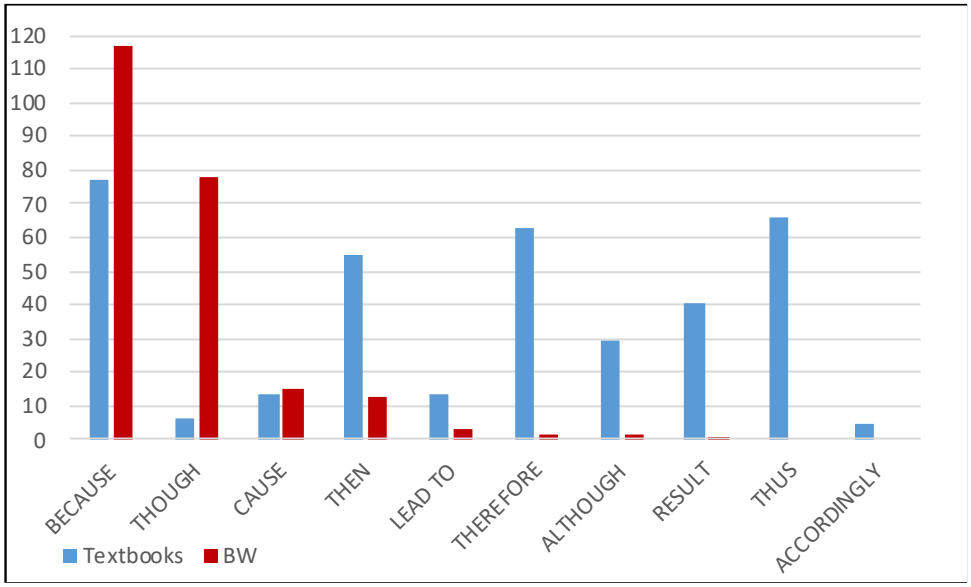


Figure 7: Causality markers normalised for 100,000 words (BW- AI-gen. text only)

Because of the low frequencies, the focus on clusters will be only on the first three words in Figure 7.

While *cause* appears proportionally with the same numbers, the semantic direction differs markedly. First of all, the cause is followed by the *to Inf.* in 6/100 cases in BW – but 12/100 in LIBB. Looking at the most frequently occurring uses of *cause*, BW use is unequally split between 'direct cause' (*will, which, readily cause)* = 40.75% of all uses, and 'indirect (hedged) cause' (*might, can, presumably cause) = 31.25%* of all uses. This is very different for LIBB where 'direct cause' (*will, that, which, therefore, triggers cause)* = 13.2% while all uses and 'indirect (hedged) cause' (*can*, *may, might, could, probably cause*) = 43.4%. A further 4/100 cases in LIBB record *not cause*, a term not found in BW.

Because has the same high-frequency tri-gram chunks: "because of the", "because of its" and "because of their". It must be highlighted that "because of" appears in almost all concordance lines

[87] It must be noted that *which* is significantly more frequent in BW, yet *that* is underused so it appears that this relative clause marker covers both forms.

in frequently in BW, while it appears in just over half the LIBB concordance lines. Furthermore, the R2 position in LIBB appears to make greater use of grammatical words as well as verbs than BW, resulting, overall, in a higher degree of formulacity in BW. This can also be seen when *because* is looked for in long repeated chunks. There is only a single 6-part chunk that occurs more than 3 times in LIBB; in BW, there are a number of 7 or 8-item chunks that occur three times.

For *though*, the trigram "even though the" is the most frequent (only) 3-gram for both – 11 (14) for BW (LIBB) – again, it is significantly overrepresented in BW (for the corpus overall) where *though* occurs 84 times while it is only used 70 times in LIBB. It appears that it is *though* which appears with a more varied use in BW in stark contrast to LIBB which has fewer (near-collocates) despite its similar absolute frequency.

2.3.2. Divergent verb usage

The most frequent prime verb forms show a clear preference for past perfect *had*, namely *had been* while LIBB overwhelmingly uses the descriptive present tense with *is* / are and the simple past *was*. Roughly, this appears to reflect that Beta Writer reports past results, whereas the textbooks detail ongoing processes and report on results obtained directly.

Table 5 provides suggestions as to what formulations the textbooks have a preference for where BW uses *enhance** or *indicate**. To do so, the most frequent trigram was used as a frame, with the target-word replaced, in LIBB, with a wildcard.

	BETA WRITER		**TEXTBOOKS**
(1)	results *INDICATE* that	->	results *SUGGEST* that
(2)	which *INDICATES* that	->	which MEANS / SHOWS that
(3)	as *INDICATED* by	->	as *DEFINED / DESCRIBED / DETERMINED / SHOWN by*
(4)	to *ENHANCE the*	->	To *IMPROVE / INCREASE* the[88]
(5)	With *ENHANCED* electrochemical	->	With *IMPROVED* electrochemical
(6)	and *ENHANCING the*	->	and *IMPROVE / INCREASE* the

Table 5: LIBB alternatives to highly frequent BW verbs

Table 5 gives a good idea why these are so markedly more frequent. (1) is interesting, as it appears that the textbooks use a more circumspect form with "results suggest" – whereas *indicate* can be seen as a less-hedged form. This is in contrast to the frame *which_*_that* where the textbooks use the more direct forms. In fact, while "which implies that" occurs twice in LIBB, "which indicates that" is not recorded at all. (3) shows that *indicated* is used in a rather to general way – this particular frame in LIBB aims to be far more specific. (4) appears to be the one exception as the frame is employed for "to enhance the" in both LIBB and BW; however, like in (3) the textbooks for a more specific way of phrasing. (5) points to a very particular use – yet, even in this case, the textbooks prefer a different word. (6), finally, is like (4) only that "and enhance the" is barely used[89].

3. Discussion

This section will analyse these findings and highlight what implications can be drawn for the teaching of English for Academic Purposes (EAP) in particular.

The apparent differences are, firstly, the concrete lack of graphical display found in BW. This

[88] *to enhance the* also occurs in LIBB, yet is 1/3 less frequent than the alternatives described above.
[89] 3 times in the 1150 concordance lines using this frame.

means that one crucial route of information dissemination has not been integrated in an artificially created text that aims to be classed as a "textbook".

Secondly, *Connectors* and *resultative adjuncts* are, together with the use of *modals* for the purpose of hedging seen as typical of academic texts. It has been found, however, that BW makes use of only a narrow range of these – far narrower, in fact, than what can be seen in the human-authored textbooks. Readers would, therefore, find *though, which, and* as well as *because* being overused in BW while a significant number of other words are rare for the first two groups of words

This is mirrored for *modals.* For example, *might* in BW is typically a hedge, while in LIBB is used to indicate causation (*might cause*). *Could* is used in a way that marks it out to function almost always as a different lexical item in BW compared to LIBB.

This stands in contrast to *boosters*. Overall, these are not frequent in either set of texts. However, BW makes little use of *very,* yet significantly overuses a range of adverbs which tend to be rare in LIBB (like *fruitful, efficiently, satisfactory*). These occur widely and appear to give Beta Writer stronger positive prosodies than typically found in textbooks. This impression is strengthened by the absence of negation markers (*non-*; *not*). In fact, some of the boosters appear in BW to indicate a high level of certainty or quality whereas their use in LIBB would be more circumspect.

Moreover, the use of several of the words investigated have not only been found to be nested in a more formulaic way: crucially, what is being said is often rather general, not to say vague. Direct comparison with LIBB has shown how the same words are occurring in a much more varied but, in particular, a far more specific and poignant way to express what the human authors want to describe.

All these points reflect issues highlighted by a number of reviewers as BW was launched:

1. *Lithium Ion Batteries* resembles less a textbook; it is more one super-extended overview of the literature published on the topic
2. As the writing AI appears to have no notion of *pragmatics* or *semantics*, crucial elements that make non-fiction writing readable are missing, namely the use of connective phrases and the distinction between different orders of importance of the facts and findings given
3. Human writing and the editing of published academic writing will have a clearly defined subject focus. However, BW seems to be hyper-focussed on a number of keywords and there displays overt repetition of long and very extended chunks of words
4. Human writing and the editing of published academic writing will aim to use markers of uncertainty (hedges), avoid subjective or emotive descriptions, while also aim for a high level of precision.

Collins (2019) shows that the idea that "writing academic books by doing no more than summarising the content of lots of existing papers – creating a kind of low-level literature review" is incorrect. It can be said for (1), that this shows the format chosen is inappropriate. This is not a book – it is a condensed combined summary based on over 1,000 published research articles. Indeed, some, like Bingham (in Claburn 2019), question whether a form of targeted search, using Google Scholar or Scopus, might not do the job equally well. As for (2) and (3), the set-up ensured that the AI used only relevant base material. Yet the AI has no recorded measure to differentiate for meaning. As a result, Conrad (2019) wonders whether biases and assumptions get carried over into the end-product. Day (2019) goes one step further and points out that a human writer's skill is shown not just by what is being included but also by what is being excluded- a feature not apparent in BW. This results in, according to Meskine and Mostaghaci (2019), a text that "remains a flat collection of summarized information, with no indication to which item is more trustworthy than the rest" Roberts and Hamilton (2020, 26) explain that "authoritativeness lies on a continuum… and [authority needs to be measured]

in an academic context". Finally, our research found that the lexical choices found in BW pinpoint the divergence to human-written textbooks as in (4). A writer needs to create something that the human reader finds valuable: Meskine and Mostaghaci (2019) and Collins (2019) and others stress that writing does not happen in a vacuum. The textbooks display clear choices when the writer use more circumscription and where they are more specific; overt boosterism is avoided or employed in a strictly prescribed way (as in the case of *satisfactory* being test result label). Moreover, human writers make consistent use of negation markers. The word choices are typically well calibrated to convey the meaning and level of certainty the writers want to express. By contrast, BW provides a high level of non-specificity, not to say, vagueness.

4. What this means for Teaching

A number of the insights gleaned from this corpus-linguistic study should be of use when further AI-writing tools are created. However, it is this paper's aim to make use of the exegesis and resulting text as a kind of computer-simulation. These are issues a teacher has to look out for and aim to avoid when doing EAP instruction.

1. *Over-reliance on keywords.* It is true that a clearly defined subject will have specific keywords that appear repeated several times. As a consequence, these appear in both BW and LIBB. The danger might be, however, that they are over-used and result in a text that is too noun-heavy yet difficult to comprehend to a reader as BW shows.
2. *Formulaic chunks.* In instructing human writers, this might be less of a problem. However, there is a parallel issue, namely the use of *quotes or paraphrases*. A learner has to be aware that a) the same point should not be repeated with very similar, or the same wording and *quotes* should not simply be implanted in a text – it has to be embedded and fully fit the context it is being found in as explained in, for example, Godfrey (2013).
3. *Word classes which are typical of academic writing*. These points are based on Carter and McCarthy (2006). Beyond any vocabulary that is subject-specific, a learner needs to understand the value of varied use of complex sentence. Consequently, there should be relative clauses introduced with both *that* and *which*. There must be a wide range of hedges. These are typical modals, and a teacher might benefit from using corpora that give a good indication when *may* should be given preference to *might* or where *can, be able to, could, will* or *should* will be employed more gainfully. This research highlights the issues *BetaWriter* has when compared directly with academic writing that is specific to this genre and subject. Figure (1) is notable in demonstrating the difference between human-created and the AI-created textbook on *lithium-ion batteries.*
4. *Prosody*. This links in with (3). This investigation has highlighted that the textbooks use negation markers whereas BW does less so; use of modals is typically fitting to reflect the writer's level of certainty. These decisions have to be made by a writer and certain words reflect the underlying intentions more fitting than others. Furthermore, LIBB appears to make very sparing and targeted use of booster words – unlike BW. To use this as an object lesson what should be considered and what to be avoided, a teacher can benefit from showing how uses vary for different words by referring to concordance lines of suitable texts.
5. *Highly targeted word and phrase use*. This, in a way, expands on points three and four. Typically, an intermediate learner will know a small set of suitable forms to express what they want to say – typically verbs, adjectives and adverbs. One concrete error that BW seems to show is overuse of adverbs, adjectives and verbs that are either unsuitable or badly targeted

(*fruitful, efficient*) or which are overused (*indicated*) when a better targeted set of wider choices are available. This means that a learner must have an understanding that a thesaurus is not a mere menu from which one can pick at random. It should not be over-used and, where employed, each words' nesting should be taken into account.

6. *Evaluation and making distinctions*. As many reviewers have pointed out, as the high number of long, repeated chunks of text indicate, BW appears to be like one long list of details. On the surface of it, BW looks like a textbook; this matches the first impressions readers might get when looking at (rather than reading closely) many essays, dissertations, or theses, as these might, indeed, look like an academic text. However, inside the text every reader needs guidance. This means that a proficient writer must leave out what is less relevant; must give indication which findings are more important, more trustworthy than others. BW fails to do this. It is a tall order to achieve for a learner, but the result will reflect a high level of discourse fluency.

5. Conclusion

Socrates once said, "Improve yourself by studying the writings of other people. In this way, you can easily gain the knowledge that other men have worked very hard for." Springer might have had the same intention when teaching Beta Writer (BW), an artificial intelligence (AI), to write a textbook on lithium-ion batteries.

Bigham (in Claburn 2019) stresses that writing a suitable textbook is a difficult task even for an experienced writer. Based on our findings, future enterprises of this kind still need to improve a lot. The corpus analyses conducted in this study suggest that the readers would not be able to easily gain the knowledge conveyed by the book due to the language it employs.

This problem might be rooted in how the researchers taught the AI to write the book as *Beta Writer* can probably be more appropriately named as "AI-assisted summarising-and-structuring literature overview".

The academic writing skills taught to BW were mainly summarising and paraphrasing. Teaching academic writing is, in fact, not just about teaching summarising and paraphrasing skills. It is also about teaching the lexis and grammatical structures that are common in the register of academic writing. As a result of our analysis, we have highlighted that teaching academic writing is also about teaching how to use the keywords of a given topic in an appropriate way; and how to use direct quotes and how to paraphrase them accordingly. It is also about teaching of how to employ word classes and grammatical structures typical of academic writing. In relation to that, students also need to be made aware of the prosodic aspects of academic texts, for example, through the suitable use of negation and modals. Moreover, teaching academic writing is also about teaching the highly targeted words and phrases; not to overuse them, vary them accordingly, and use them in a form that is fitting the context. Finally, in writing an academic text, it is also important for the writer to evaluate what she has been reading; what to include and exclude in her writing. Errors apparent in Beta Writer can be used to teach teachers what to be aware of when teaching students how to write academic texts.

References

Bailey, D., Southam, A. and Costley, J. (2021), "Digital storytelling with chatbots: mapping L2 participation and perception patterns", Interactive Technology and Smart Education, Vol. 18 No. 1, pp. 85-103. https://doi.org/10.1108/ITSE-08-2020-0170

Bailin, A. (1987). Artificial intelligence and computer-assisted language instruction: A perspective. *Calico Journal*, 5(3), 25–45. https://doi.org/10.1558/cj.v5i3.25-45

Beta Writer (2019). *Lithium-Ion Batteries. A Machine-Generated Summary of Current Research.* Cham: Springer International Publishing.
Carter, R. and McCarthy, M. (2006). *Cambridge Grammar of English*. Cambridge: CUP.
Claburn, T. (2019). Want to learn about lithium-ion batteries? An AI has written a tedious book on the subject. *The Register*. https://www.theregister.com/2019/04/08/ai_generated_book/
Collins, H. (2019). Death of the author? AI generated books and the production of scientific knowledge. https://blogs.lse.ac.uk/impactofsocialsciences/2019/05/09/death-of-the-author-ai-generated-books-and-the-production-of-scientific-knowledge/
Conrad, L. Y. (2019). The Robots are Writing: Will Machine-Generated Books Accelerate our Consumption of Scholarly Literature? *The Scholarly Kitchen*. https://scholarlykitchen.sspnet.org/2019/06/25/the-robots-are-writing-will-machine-generated-books-accelerate-our-consumption-of-scholarly-literature/
Day, C. (2019). Here come the robot authors. *Physics Today* 72, 6-8; https://doi:10.1063/PT.3.4213
D'Amour, A. K. Heller, D. Moldovan, et al. (2020). Underspecification Presents Challenges for Credibility in Modern Machine Learning. https://arxiv.org/pdf/2011.03395.pdf
El Shazly, R. (2021). Effects of artificial intelligence on English speaking anxiety and speaking performance: A case study. *Expert Systems,* 38 (3), e12667.
Gallacher, A., Thompson, A., & Howarth, M. (2018). "My robot is an idiot!"–Students' perceptions of AI in the L2 classroom. Future-proof CALL: language learning as exploration and encounters–short papers from EUROCALL 2018, 70.
Ghoneim, N. M. M. & Elghotmy, H. E. A. (2021). Using an Artificial Intelligence Based Program to Enhance Primary Stage Pupils' EFL Listening Skills. *Sohag University Journal of Education* 83, 1–32.
Godfrey, J. (2018). *How to use your reading in your essays*. Macmillan International Higher Education.
Haristiani, N. (2019). Artificial Intelligence (AI) chatbot as language learning medium: An inquiry. *Journal of Physics: Conference Series,* 1387 (1). IOP Publishing, 012020.
Hoey, M. (2005). *Lexical Priming*. A new theory of words and language. London: Routledge.
Holland, V. M., Maisano, R., Alderks, C., & Martin, J. (1993). Parsers in tutors: What are they good for? *Calico Journal,* 11(1), 28–46. https://doi.org/10.1558/cj.v11i1.28-46
Hyland, K. and Tse, P. (2007). Is there an "Academic Vocabulary"? TESOL Quarterly 41 (2), 235–253.
Jehle, F. (1987). A free-form dialog program in Spanish. CALICO Journal, 5(2), 11-22.
Köse, U., & Arslan, A. (2015). E-Learning experience with artificial intelligence supported software: An international application on English language courses. GLOKALde, 1(3), 61–75.
Last, R. (1989). *Artificial Intelligence Techniques in Language Learning.* New York: Ellis Horwood.
Lu, X. (2018). Natural Language Processing and Intelligent Computer-Assisted Language Learning (ICALL). The TESOL Encyclopedia of English Language Teaching, 1–6. https://doi.org/10.1002/9781118784235.eelt0422
Manning C.D. (2015) Computational linguistics and deep learning. *Computational Linguist* 41(4):701–707. https://doi.org/10.1162/COLI_a_00239
Meskine, H and Mostaghaci, B. (2019). Beta's Draft: A Human Review of a Machine-generated Book. *Advanced Science News.* https://www.advancedsciencenews.com/betas-draft-a-human-review-of-a-machine-generated-book/
Nagata, N. (1996). Computer vs. workbook instruction in second language acquisition. *Calico Journal,* 14(1), 53-75.
O'Brien, P. (1993). eL: AI in CALL. In M. Yazdani (ed.), *Multilingual Multimedia. Bridging the Language Barrier with Intelligent Systems,* 85–139.
Pace-Sigge, M. (2018). *Spreading activation, lexical priming and the semantic web: early psycholinguistic theories, corpus linguistics and AI applications*. Abingdon: Palgrave Macmillan.
Rayson, P. (2016) *Log-likelihood calculator*. Downloadable from http://ucrel.lancs.ac.uk/llwizard.html
Rieland, R. (2017). Is artificial intelligence the key to personalized education? Smithsonian Magazine. https://www.smithsonianmag.com/innovation/artificial-intelligence-key-personalized-education-180963172/
Roberts, J. Q., & Hamilton, C. (2020). *Reading at University: How to Improve Your Focus and be More Critical*. Red Globe Press.
Salaberry, M. R. (1996). A theoretical foundation for the development of pedagogical tasks in computer mediated communication. CALICO Journal, 14 (1), 5–34. https://doi.org/10.1558/cj.v14i1.5-34
Scott, M. (2020). WordSmith Tools version 8, Stroud: Lexical Analysis Software.
Seidel, R. J., & Park, O. C. (1994). An historical perspective and a model for evaluation of intelligent tutoring systems. *Journal of Educational Computing Research*, 10(2), 103-128.
Self, J. (1998). The defining characteristics of intelligent tutoring systems research: ITSs care, precisely. *International Journal of Artificial Intelligence in Education (*IJAIED), 10, 350–364.
Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, 106(2), 331–347. https://doi.org/10.1037/a0034752
Sumakul, D. T. (2019). *When robots enter the classrooms: Implications for teachers*. In *Embedding artificial intelligence (AI) in education policy and practice for Southeast Asia* (pp. 36-42). SEAMEO

SEAMOLEC.
Sumakul, D. T., Hamied, F. A., & Sukyadi, D. (2021). Students' Perceptions of the use of AI in a writing class. *The 67th TEFLIN International Virtual Conference & The 9th ICOELT 2021*. Atlantis Press.
Wolff, D. (1993). New Technologies for Foreign Language Teaching. In Foreign Language Learning and the Use of New Technologies. Conference Proceedings. London 1993. Brussels: Bureau Lingua.
Yoshio, M., Brodd, R. J., & Kozawa, A. (2009). *Lithium-ion batteries*. New York: Springer.
Zawacki-Richter, O., Marín, V., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? International Journal of Educational Technology in Higher Education. https://doi.org/10.1186/s41239-019-0171-0

(all URLs last accessed 01/April/2022)