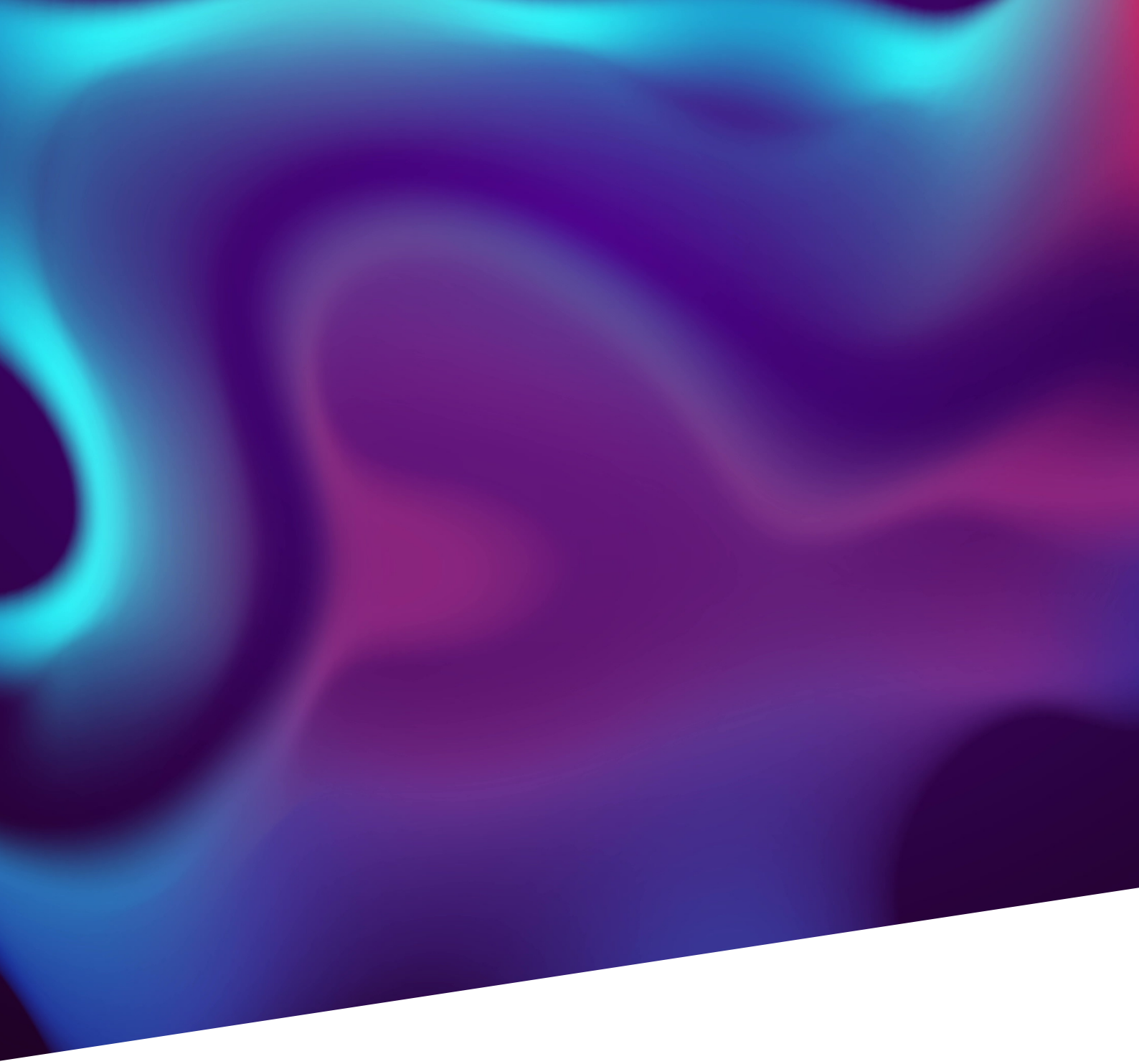

3RD INTERNATIONAL CONFERENCE ON
'LANGUAGE IN THE HUMAN-MACHINE ERA'

BOOK OF ABSTRACTS

15-16 MAY 2023 UNIVERSITY OF GRONINGEN

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

LANGUAGE IN THE
HUMAN-MACHINE ERA

3rd International Conference on 'Language in the Human-Machine Era'

Book of Abstracts

15-16 May 2023 University of Groningen

LANGUAGE IN THE HUMAN-MACHINE ERA

This publication is based upon work from COST Action 19102 'Language in the Human-Machine Era', supported by COST (European Cooperation in Science and Technology).

COST (European Cooperation in Science and Technology) is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career and innovation.

www.cost.eu

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

Preface

We are glad to introduce this book of abstracts of the third annual international conference "Language in the Human-Machine Era" (LITHME). The LITHME network funded by European Cooperation in Science in Technology (COST) creates opportunities for sustainable dialogue among communities in language and technology research, and the LITHME conference is a venue where researchers associated with communities that rarely interact with each other can engage in meaningful exchange. In 2023, the conference was organised at the University of Groningen, Campus Fryslân, in Leeuwarden (Netherlands) from 15th to 16th May, and hosted by the Speech Technology Lab.

The LITHME conference program features two plenary talks, one rapid networking session, one advice panel from industry professionals to academics, and 40 talks grouped into 14 sessions and two parallel streams that interested people could attend for free online or in person. The sessions covered technological and linguistic aspects of topics such as translation, interpretation, large language models, language variation, language rights, language and power, low-resource and minority languages, language learning, language vitality and natural language processing.

We thank our keynote speakers Daan Van Esch (Google Research) and Jieun Kiaer (University of Oxford) for their inspiring and informative talks on 'Building Language Technologies across the World's Languages' and 'Language with AI: A Linguist's response to ChatGPT'.

LITHME conference is a result of dedication and hard work of our community and our supporters. We thank all LITHME members who received, reviewed, and evaluated the submissions, processed invitations, and last-minute changes in the program. The hybrid nature of the event increased the workload of the local organisers. We thank our local organisers' team led by Matt Coler (Speech Technology Lab) and the SpeechTech Summer School that took place in parallel for the poster session.

Finally, it was our great pleasure to welcome our onsite and remote participants! We hope that you enjoyed the conference and hope to see you at the 4th LITHME conference in 2024!

For more information about LITHME's activities and interests, readers are encouraged to look at our website (https://lithme.eu), our open access forecast report (https://doi.org/10.17011/jyx/reports/20210518/1) and our professionally commissioned animations (https://lithme.eu/animations).

LITHME Conference Program Chairs

Sviatlana Höhn, Britta Schneider, Dave Sayers, Rui Sousa Silva

Learning

1 Using Educational Digital Storytelling with Multilingual and Multicultural Students: Implications for Enhancing Learners' Writing Skills

Eleni Meletiadou, London metropolitan University

According to research, Educational Digital Storytelling (EDS) is a pioneering, learning-oriented, technology-enhanced approach that allows students to develop a wide range of academic and professional skills (Yildiz Durak, 2018). Barrett (2019) claims that EDS combines four student-centered learning strategies: student engagement, reflection for deep learning, project-based learning, and the effective integration of technology into instruction. EDS is also renowned because it promotes reflection (Jamissen et al., 2017) and collaboration. The current study explored the use of EDS with undergraduate international business school students to develop multilingual and multicultural students' writing skills and enhance their motivation towards learning in the post-Covid era. 50 first-year students were randomly asked to participate in this study in terms of an undergraduate module (Digital Business Management and Emerging Technologies). These students were asked to tell their own digital stories using either blogs, websites, AR, or VR and were engaged in interactive seminars for a whole academic semester. EDS provides an opportunity for students to solve problems and develop their digital skills while becoming more confident with technology through experimentation. This study intended to investigate how to use EDS activities to facilitate students' digital literacy in HE. The overall aim of the current study was to explore the potential benefits of EDS in business schools and ultimately to promote its use with business and management students as it fosters strategic thinking, self-reflection, teamwork, and the development of digital and professional skills. The results will hopefully advance the understanding of digital literacy development through EDS activities in multicultural and multilingual HE classrooms. The findings of this study showed that EDS is a promising approach for improving writing skills. Considering the need for immediate change in multi-modal writing skills in current HEI classes, the findings of this study can indicate ways in which educators can implement this technology-enhanced learning method in terms of which students can exchange ideas and improve their writing skills using both verbal and non-verbal elements. Educators in 21st-century HEIs can benefit from this highly interactive approach to help students develop their academic skills while they engage in the development of engaging stories that allow them to put the theory they have learnt into practice. Learners feel flattered because they have a real audience and spend considerable time studying theories and applying them, creating real-life scenarios which may help them to develop their professional skills. Therefore, in terms of the EDS-integrated instruction, students find the development of academic skills, ie, writing more meaningfully and engagingly, and are willing to reflect on their work and that of their peers as they strive to improve each other's skills to achieve their final learning goals. However, educators need to be cautious when using this exciting new technique. Students face considerable problems using this tool when they have limited access to the internet or technology. Some of them, especially mature students, have limited knowledge and experience in using digital skills in their everyday life. Therefore, educators should provide necessary training and continuous support before and while using it to avoid any kind of discrimination. In conclusion, EDS seems to be a viable tool for HEI lecturers who would like to enrich traditional undergraduate courses and allow students to have access to and become proficient in 21st- century multi-modal literacies.

Keywords: Language, Virtual Reality, Augmented Reality, Educational Digital-Storytelling, Human-Machine interfaces

Anderson, J., Chung, Y. C., & Macleroy, V. (2018). Creative and critical approaches to language learning and digital technology: Findings from a multilingual digital storytelling project. *Language and Education*, *32*(3), 195-211.

Balaman, S. (2018). Digital storytelling: A multimodal narrative writing genre. *Journal of Language and Linguistic Studies*, *14*(3), 202-212.

Barrett, A. K. (2019). Digital storytelling: Using new technology affordances to organize during high uncertainty. *Narrative Inquiry*, *29*(1), 213-243.

Chiang, M. H. (2020). Exploring the effects of digital storytelling: A case study of adult L2 writers in Taiwan. *IAFOR Journal of Education*, *8*(1), 65-82.

Hava, K. (2021). Exploring the role of digital storytelling in student motivation and satisfaction in EFL education. *Computer Assisted Language Learning, 34*(7), 958-978.

Jamissen, G., Hardy, P., Nordkvelle, Y., & Pleasants, H. (2017). Digital storytelling in higher education. *International perspectives*.

Liu, K. P., Tai, S. J. D., & Liu, C. C. (2018). Enhancing language learning through creation: The effect of digital storytelling on student learning motivation and performance in a school English course. *Educational Technology Research and Development*, *66*(4), 913-935.

McLellan, H. (2008). Digital storytelling: Expanding media possibilities for learning. *Educational Technology*, 18-21.

Meletiadou, E. (2021). Using Padlets as e-portfolios to develop undergraduate students' writing skills and motivation. *IAFOR Journal of Undergraduate Education*, *9*(5), 67-83.

Niemi, H., Niu, S., Vivitsou, M., & Li, B. (2018). Digital storytelling for twenty-first-century competencies with math literacy and student engagement in China and Finland. *Contemporary Educational Technology*, *9*(4), 331-353.

Özüdoğru, G., & Çakir, H. (2020). An investigation into the opinions of pre-service teachers toward uses of digital storytelling in literacy education. *Participatory Educational Research*, *7*(1), 242-256.

Yildiz Durak, H. (2018). Digital story design activities used for teaching programming effect on learning of programming concepts, programming self-efficacy, and participation and analysis of student experiences. *Journal of Computer Assisted Learning, 34*(6), 740-752.

The current project won the British Academy of Management Education Practice Award and showcases how EDS (technology-enhanced pedagogical approach) can be used to support international multilingual students develop their language and content (Business Management) skills at a Business School that promotes multilingual and multicultural awareness and places an emphasis on the development of digital skills. This presentation related with your language learning and teaching working group (I am a member of this group) which explores how VR and AR can be used to promote language learning in HE supporting international multilingual and multicultural students globally.

2 Combining different approaches to monitor the linguistic development of Basque university students.

Jose Mari Arriola, UPV/EHU

Mikel Iruskieta, UPV/EHU; Ekain Arrieta, UPV/EHU

Basque multilingual education systems require to achieve the appropriate CEFR level at different education levels by the present Decree-Law, but there are no data-evidences nor reliable and massive language analysis. For example, Osinalde and Iruskieta (2022) analyzed a small sample of written text to describe the most notable errors in two CEFR levels. Language Technologies can be an important step forward in providing data-evidences in language learning and teaching. We are convinced that they are of strategic importance for the revitalization and the future of languages, especially in the case of less-resourced languages such as Basque, because everyone can understand Basque using machine translation (text-to-text) or speech recognizers and machine translation (speech-to-text and translated text-to-speech). Language Technologies can help towards an empirical validation of CEFR required levels in the multilingual educational systems of the Basque Country. In this context, this project aims to answer a fundamental question that the Basque trilingual educational systems have regarding the entry and the exit language profiles of students. Various initiatives have been launched to promote and analyze Basque in three Faculties of Education at the UPV/EHU, but the high number of students makes it difficult and expensive to diagnose and monitor how the language is developing in this context or to ensure that an adequate language level is reached at each academic stage. The university requires students to have a B2 CEFR level in Basque in the 1st year of the grade and C1 CEFR level after finishing the 4th year of university studies by the present Decree-Law, but data-evidences are needed to evaluate the language acquisition level in this trilingual educational environment. Thus, this project seeks to design automatic tools that help students develop academic literacy in Basque by considering their needs. For this purpose, two objectives have been established: (1) to carry out a detailed study of linguistic discursive characteristics based on various corpora written in Basque in the Educational field and (2) to create language analyzers for Basque and integrate them into the CLARIN infrastructure alongside previously created tools for other languages. We will present the work carried out on a sample of 480 student essays in B1, B2, C1, and C2 language level tests (Arrieta et al., In press). To do so, we have used different machine learning techniques (SVM) and language models to predict the level of the student's writings. We obtain an accuracy of 93% when predicting the CEFR level (B1-C2) of a written essay. Regarding the binary prediction of whether an essay reaches level C1, we obtain an accuracy of 97%; and the average accuracy when predicting whether an essay would pass the writing test of a specific level is 78%. We will also present some ways that we have identified to improve these results. Similar to Hnatkowska and Wawrzyniak (2022) we obtained the best results with SVM, but we are considering other techniques such as RNN classifiers trained on complexity contours similar to Kerz et al. (2021). Moreover, we are working on developing reusable tools and corpus (following FAIR principles), such as Basque integration in the CTAP system (Chen & Meurers, 2016). In the foreseeable future, we would like to create for Basque other automatic systems.

Keywords: **Automatic text classification**, **Tools for literacy**, **Multilingual systems**

Arrieta, E., Odriozola, I., Arregi, X., Iruskieta, M. (In press). HABE-IXA euskarazko idazmen-proben corpuseko idazlanen mailakatze automatikoa. *eHizpide*.

Chen, X., & Meurers, D. (2016). CTAP: A web-based tool supporting automatic complexity analysis. *In Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pages 113–119, Osaka, Japan, December 11-17.

Hnatkowska, B., & Wawrzyniak, D. (2022). Proficiency Level Classification of Foreign Language Learners Using Machine Learning Algorithms and Multilingual Models. *In Computational Collective Intelligence: 14th International Conference*, ICCCI 2022, Hammamet, Tunisia, September 28–30, 2022, Proceedings (pp. 261-271). Cham: *Springer International Publishing*.

Kerz, E., Wiechmann, D., Qiao, Y., Tseng, E., & Ströbel, M. (2021). Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs. *In Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 199-209).

Osinalde, M., & Iruskieta, M. (2022). Hizkuntza ikasleen testu corpus etiketatuaren analisia eta interpretazioa B2 eta C1 mailetan [Analysis and interpretation of language students' annotated text corpus at B2 and C1 levels]. *eHizpide 100*. https://doi.org/10.54512/HLFA9295.

3

Augmented Reality: an attractive way to teach languages for the new pre-service teachers' generation?

Eve Lejot, University of Luxembourg

Corine Philippart, University of Luxembourg

The major educational concern of the 21st century is the integration of digital technologies in the classrooms. However, this shift must not conceal the fact that new technologies are not an end in themselves. They are first and foremost a means, one among others, which aims to bring added value to a task that would not have been possible otherwise (Noben, 2022: 45-47). This question of the effective use of technology, or in other words, the need to generate a "pedagogical added value of digital technology", has already appeared with the first technological breakthroughs (Punar Özçelik et al., 2022: 134). In retrospect, with the first films, Edison promised in the 1910s to revolutionize education. Since then, other technological developments have followed with the telephone, radio, television, video, and finally the computer and the Internet in the 1980s (Karsenti, 2004: 7). All of them have to a certain extent found their uses in classrooms over time. The democratization of another technology, i.e. the smartphone, have also greatly influenced the educational landscapes. With its small size, its limited weight as well as its computing capabilities, the smartphone has made it possible to integrate new tools into educational practices, generating great enthusiasm on both sides of the classroom. In particular, the new generations of teachers tend to be more favorable to the use of ICT in their teaching (Olafare et al. 2018). This is imputed to the fact that students are "digital natives", as they are considered to have been educated in a digital world (Prensky, 2001). In this communication, we question the assumption that the new generations, and more precisely the pre-service teachers, are, by virtue of their life experience, digital natives. Challenging this assumption will lead us to launch a case study, conducted on teachers in training in multilingual environments. Given that the technical proficiency in Augmented Reality is one of the most significant issues of this technology, our focus will be on the introduction of Augmented Reality to pre-service high school teachers (Belda-Medina & Calvo-Ferrer, 2022). Our presentation will be structured in three points. First, we will start with a theoretical discussion on the notion of interactive technology and especially, of Augmented Reality. We will then detail the sequence of the case study design. The description of the case study will allow us to examine the representations of these pre-service teachers of Augmented Reality as part of the methodology they would consider using in the classroom.

Keywords: Augmented Reality (AR), Case study, Teachers in training, Representations, Digital teaching

Belda-Medina, J., & Calvo-Ferrer, J. R. (2022). Integrating augmented reality in language learning: pre-service teachers' digital competence and attitudes through the TPACK framework. *Education and Information Technologies*, *27*(9), 12123–12146. https://doi.org/10.1007/s10639-022-11123-3

Prensky, M. (2001). Digital Natives, Digital Immigrants Part 2: Do They Really Think Differently? *On the Horizon*.

Karsenti, T. (2004). Pourquoi une revue scientifique internationale portant sur l'intégration des TIC en pédagogie universitaire?. *Revue internationale des technologies en pédagogie universitaire*, *1*(1"), 7-8. https://doi.org/10.18162/ritpu.2004.16

Olafare, F. O., Adeyanju, L. O., & Fakorede, S. O. A. (2018). Colleges of Education Lecturers Attitude Towards the Use of Information and Communication Technology in Nigeria. *MOJES: Malaysian Online Journal of Educational Sciences*, *5*(4), 1-12.

Noben, N. (2022). Les plus-values pédagogiques liées à l'intégration du numérique: les représentations d'étudiants du master en sciences de l'éducation de l'Université de Liège. *Revue internationale des technologies en pédagogie universitaire*, *19*(3), 44-59. https://doi.org/10.18162/ritpu-2022-v19n3-03

Punar Özçelik, N., Yangin Eksi, G., & Baturay, M. H. (2022). Augmented Reality (AR) in Language Learning: A Principled Review of 2017-2021. *Participatory Educational Research. 9*(4). 131-152. http://dx.doi.org/10.17275/per.22.83.9.4

4 "Will our students be dispensed from thinking?" – Challenges for students and teachers when AI meets language learning

Elena Gallo, LMU Munich

In the history of humanity, we have faced some major technical revolutions and very often they were perceived as end-of-the-world scenarios. Now that we are in the era of machine text production and lesson creation using AI, even teachers will have to face the fear that machines can outperform them. Are we now at a similar crucial point in human history? And are these new developments positive? Yes and no. While many language teachers show enormous concerns and even fear of these new tools, with some of them being overwhelmed by the idea of using more than an overhead projector, some researchers show the extensive and complex use of technology in language work (Sayers et al., 2021) and suggest that there is a profound mutual influence and an "increasing permeability "of pedagogies of learning and technology (Reinders & White, 2016), and other actors in the field of training and competence development see no danger for learning as a life-long process because it "will become a mainstay of training and education" and "pedagogy […] will remain a critical part of digital learning". (Acolad Team, 2023). Undoubtedly there are many useful aspects in the deployment of AI tools in foreign/second language (L2) learning and work, but they also hide some threats and challenges (European Parliament, 2022). I think the debate promoted by LITHME is pivotal to exploring and reflecting on this epochal issue in order to be better prepared as L2 teachers of young adults and to prepare them for their future life. Numerous questions arise around this issue, such as how we can manage this emerging phenomenon without falling back into an end-of-the-world scenario? How can we equip teachers and especially L2 teachers in this endeavour? How can we best deal with teachers' resistance and attitudes towards critical innovations (Gallo, 2012; Pop, 2010) and still maintain a positive affective-emotional balance (Gallo, 2016)? As a consequence of these emerging innovations, L2 teachers will surely and more than ever be in need of entering into dialogue with emerging technology developments and reviewing their attitudes and their teaching practices. However, little empirical work in L2 learning and teaching has been done to investigate emerging technologies in L2 learning (Klimova et al., 2023) and how language teachers may cope with this issue. By combining various language learning tasks for learning Italian and instruments (self-reports, questionnaires, interviews, students group discussion, and teachers' logs), this empirical research study investigates how university language teachers have to rethink the tasks they give to their students if they want to promote critical thinking in language learning, and what the students' reactions are. Since the impact of the new technologies on L2 instruction is assumed to be positive (Tuomi, 2018), this study aims to prepare language teachers and learners as well for these emerging innovations and to understand the teacher's scope of action and the implications for L2 teacher education. (Please note that I would be flexible if the WG Chairs would prefer me to participate in a roundtable discussion.)

Keywords: Emerging technologies, L2 learning, L2 teachers

Acolad Team (2023). How do technologies affect the future of the language service business? *acolad.* https://blog.acolad.com/how-do-technologies-affect-the-future-of-the-language-service-business

European Parliament (2022). Artificial intelligence: threats and opportunities. https://www.europarl.europa.eu/news/en/headlines/society/20200918STO87404/artificial-intelligence-threats-and-opportunities

Gallo, E. (2012*). Teacher professional development. How university language teachers approach their own professional development.* (Publication no. 17688) [Doctoral dissertation, Ludwig-Maximilians-Universität]. https://edoc.ub.uni-muenchen.de/17688/

Gallo, E. & M.G. Tassinari, (2017). "Positive feelings about my work: I needed it!' Emotions and emotion self-regulation in language teachers". *Apples: Journal of Applied Language Studies*, 11(2), pp. 55-84. https://doi.org/10.17011/apples/urn.201708233539

Klimova, B. Pikhart, M., Polakova, P., Cerna, M., Yayilgan, S.Y., Shaikh, S. (2023). A Systematic Review on the Use of Emerging Technologies in Teaching English as an Applied Language at the University Level. *Systems, 11*(42). https://doi.org/10.3390/systems11010042

Pop, A. (2010). The impact of the new technologies in foreign language instruction our experience. *Procedia - Social and Behavioral Sciences*, *2*(2), 1185–1189. https://doi.org/10.1016/j.sbspro.2010.03.169

Reinders, H. & White, C. (2016). 20 years of autonomy and technology: How far have we come and where to next? *Language Learning & Technology*, *20*(2). 143–154. http://dx.doi.org/10125/44466

Sayers, D., Sousa-Silva, R., Höhn, S., Ahmedi, L., Allkivi-Metsoja, K., Anastasiou, D., ... & Yayilgan, S. Y. (2021). The Dawn of the Human-Machine Era: A forecast of new and emerging language technologies. [Report for EU COST Action CA19102 'Language In The Human-Machine Era']. www.lithme.eu.

Tuomi, I. (2018). The Impact of Artificial Intelligence on Learning, Teaching, and Education. Policies for the future, Eds. Cabrera, M., Vuorikari, R & Punie, Y., EUR 29442 EN, *Publications Office of the European Union, Luxembourg.* https://publications.jrc.ec.europa.eu/repository/handle/JRC113226

Variation

5 The performativity of dialect recognition technology in asylum determination

Jasper van der Kist, European University Viadrina

Refugee status determination is increasingly being shaped through and within technologies. A notable development in recent years is that some asylum authorities in Europe have started experimenting with language and dialect recognition technologies to determine the country of origin of asylum seekers. Prime example is the Dialect Identification Assistant (DIAS), developed by the German Federal Office for Migration and Refugees (BAMF). Germany has been using language analysis to verify people's origin since the 1990s, but is now automating this process using speech recognition algorithms that allow officials to identify which country undocumented migrants come from based on their dialect. While there is increasing attention in the literature on the implications of digital language technologies (Dippold et al. 2002), this paper argues that more research should be conducted on the recent appropriations of the same (or similar) technologies by government institutions, especially in the context of migration management. Moreover, it argues that Science and Technology Studies (STS), and Actor-Network Theory (ANT) in particular, can revitalise interdisciplinary debates about the performative and political dimensions of language in the human-machine era. Within language studies, the notion of 'performativity' has been adopted to argue that "identities are formed in the linguistic performance rather than pregiven" (Austin & Warnock, 1964; Pennycook, 2004). In addition, (critical) migration studies have also long argued that the identity of a migrant is socially constructed rather than a starting point or a pre-existing reality that can simply be revealed and acted upon (Huysmans, 2006). Moving beyond epistemological registers, this paper draws from material-semiotic conceptions of performativity (Callon, 2007; Law, 2004; Mackenzie, 2006; Scheel et al., 2019). It shows how the Dialect Identification Assistant (DIAS) helps to bring into being the very identities it is meant to identify because they are part of wider sociotechnical arrangements – including, amongst others, material artefacts and devices, data infrastructures, (geo)linguistic theories and models, communities of (computer-)linguistic practice, as well as a wider political economy of knowledge production. This analysis provides the basis for a renewed discussion of how language technology represents a new form of technopolitics (Schneider, 2022). It raises the question of whether a critique of (computer-)linguistics as 'biased' or 'inaccurate' offers a sufficient basis for critique. Building on Annemarie Mol's (2002) notion of 'ontological politics', it argues that these language technologies are not the objective or neutral representations they often claimed to be, nor simply 'ideological' in a reductive sense. If linguistic identity is a fragile and disputed accomplishment, invested with political and institutional agendas as well as commercial interests, and without sufficient attention to the interests of asylum seekers, then the politically salient question is how can these technological innovations be renegotiated and democratised?

Keywords: Language recognition technology, Science and technology studies, Asylum determination, Computational linguistics, Dialect analysis

6 Human-machine communication strategies in today's Esperanto community of practice

Federico Gobbo, Universiteit van Amsterdam

Although Esperanto was published in 1887 and since then it has been a living language, in general, it receives little attention in its community of practice, with few notable exceptions (Fians, 2021, Fiedler and Brosch 2022). People who become supporters of Esperanto may be early adopters of the new technologies of their respective generations, from early radio amateurs at the beginning of the 20th c. (Garvía 2015) to free software evangelists at the beginning of the 21st c. (Gobbo 2004). This contribution illustrates the strategies of the Esperanto community of practice in human-machine communication in the aftermath of the Covid year 2020 and how the worldwide scenario change (re)shapes its self-perceived (emic) identity in contrast with a more objective viewpoint (etic) thanks to the relatively exceptional exposure of Esperanto online thanks to the Duolingo course but not only (Gobbo 2021). Such strategies are compared to other less-resourced and minoritized languages, so to draw best practices that have broader applicability beyond the case study of Esperanto. The contribution may be framed within LITHME WGs 4, 5, and 6.

Keywords: Language ideology, Post-covid, Esperanto, Lesser-resourced languages

Fians, G. (2021). *Esperanto revolutionaries and geeks: Language politics, digital media and the making of an international community*. Cham: Palgrave Macmillan.

Fiedler, S. and Brosch C. R. (2022). *Esperanto: Lingua Franca and Language Community*. John Benjamins.

Soto, R. G. (2015). *Esperanto and its rivals: The struggle for an international language*. University of Pennsylvania Press.

Gobbo, F. (2021). Coolification and Language Vitality: The Case of Esperanto. Languages, 6(2), 93. https://doi.org/10.3390/languages6020093

Gobbo, F. (2004). Linukso kiel malterorisma vojo al informadiko. *La Gazeto*, 112, 16--23.

GPT

7 Reinforcing Ideologies of Denotational Literalism: ChatGPT as an an Epistemic Consumption Object

Michael Castelle, University of Warwick

Generative Pre-Trained (GPT) neural network architectures (Radford et al., 2018)— a form of multi-layered "decoder" neural networks which compose the second half of the "encoder-decoder" Transformer model (Vaswani et al., 2017)— have recently attained widespread popularity with the public release of an interactive 'ChatGPT' model developed by the San Francisco startup OpenAI (2022). Because the GPT-style mechanism for production of text-sentences—with its inert weight parameters, semi-arbitrarily tokenized orthography, and parallelized, feed-forward linear and non-linear computation—differs so much from traditional cognitive frameworks for generative syntax and semantics, the question remains as to what ideology of language (Schieffelin et al., 1998) such contemporary model-artifacts induce, not just within their hundreds of billions of weight parameters but in their now-ethnographically observable conditions of widespread everyday use. In this paper, I will propose that in considering ChatGPT as a site of empirical investigation we can draw from anthropological and sociological theories of consumption, including Detlev Zwick and Nikhilesh Dholakia's epistemic consumption object, which differs from traditional conceptions of consumption in that they "reveal themselves progressively through interaction, observation, use, examination, and evaluation" and become "a continuous knowledge product for consumers" (Zwick & Dholakia, 2006). Specific attention will be paid to the sociotechnical mechanism by which ChatGPT differs from previous generative transformer models, namely the use of so-called reinforcement learning with human feedback (RLHF) (Stiennon et al., 2022, Ouyang et al., 2022), in which human-evaluated scalar rankings of model outputs are used to train a secondary reward model. We will closely analyse the instructions given to these human evaluators in order to argue that an important effect of this procedure is to overtly enforce the reproduction of an ideology of what Silverstein (2023) calls "denotational literalism". Under this doctrine, instead of empirically observing communicative events in discursive social reality, we (1) attend, much like the original Transformer architectures, primarily to modally predicating-and-referring text-sentences, despite the complexity of both naturally occurring language interaction and the artifactual corpora on which GPT models are trained; and (2) assume that anyone who "knows" the language at hand have comparable "literal" understandings of any word or expression—an ideology overtly formalized in the Transformer architecture, where for any given processual instantiation of a pre-trained model, each distinct subword is consistently represented by, e.g., the same 2048-element numerical vector (Radford et al., 2019). Despite these limitations, I will show how instances of everyday epistemic-consumptive ChatGPT use manage to transcend the designers' limitations of this doctrine of literal denotation.

Keywords: Generative AI, Language ideologies, Consumption, Linguistic anthropology

OpenAI. (https://openai.com/) 2022, November 30.

ChatGPT: Optimizing Language Models for Dialogue. (https://online-chatgpt.com/)

OpenAI. (https://openai.com/blog/chatgpt/)

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022).

Training language models to follow instructions with human feedback (arXiv:2203.02155). *arXiv*. https://doi.org/10.48550/arXiv.2203.02155

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & others. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 9.

Schieffelin, B. B., Woolard, K. A., & Kroskrity, P. V. (Eds.). (1998). *Language Ideologies: Practice and Theory*. (Vol. 16) Oxford University Press.

Silverstein, M. (2022). *Language in Culture: Lectures on the Social Semiotics of Language*. Cambridge University Press.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., ... & Christiano, P. F. (2022). Learning to summarize from human feedback (arXiv:2009.01325). *arXiv*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. https://arxiv.org/abs/1706.03762v5

Zwick, D., & Dholakia, N. (2006). The Epistemic Consumption Object and Postsocial Consumption: Expanding Consumer-Object Theory in Consumer Research. *Consumption Markets & Culture, 9*(1), 17–43. https://doi.org/10.1080/10253860500481452

8 Processing and depicting novel compounds: a comparison of compositional processing in children and generative machines

Prajit Dhar, University of Groningen

Janis Pagel, University of Cologne; Lonneke van der Plas, Idiap Research Institute, Martigny, Switzerland

Compounds are compositions of two or more known lexical entities that function as a single lexical unit, for example, apple pie, fake news, mother-in-law, etc. Compounding represents one of the most productive word formation types in many languages (Baroni et al., 2002), and novel concepts in language are often denoted by means of novel compounds. Compound interpretation is an interesting process, because the relation between the two components is typically left underspecified. Baby oil is oil used for babies whereas olive oil is oil made out of olives (Shwartz & Waterson, 2018). These relations are inferred by humans based on world knowledge or conventions, but can be challenging for machines, and all the more so when processing newly created compounds. Recently, generative models such as Craiyon (formally DALL-E Mini: Dayma & Cuenca, 2023) have attracted a lot of attention, also with respect to the creative output it is able to generate, such as an image of an astronaut on a horse in space, given a text prompt. In our experiments, we aim to test these new technologies on a challenging problem that requires world knowledge and compositional processing, and is multimodal in nature, in order to shine a critical light on these new technologies. We would like to find answers to questions such as: When DALL-E is presented with a novel compound, is it able to combine the two known concepts in an intuitive way? Does the way DALL-E represent such compounds resemble what humans would make of it? How do people judge the output of such AI tools? In search of answers to these questions, we conducted an experiment with human participants (mainly children) during the Zpannend Zernike event (https://zpannendzernike.nl/activiteit/samen-stelling-samenstelling/) in Groningen. During the experiment, children were asked to draw the combination of two words (i.e., a compound), which were randomly generated based on a dice roll with pictograms on it. The compound is guaranteed to be a new concept as it does not exist in a dictionary. Once the child had completed the drawing, they were asked to evaluate what an AI-based image generation tool "drew" for the same compound. In the end we received 43 drawings from the participants. Additionally we asked the participants to: 1) select the automatically generated image that best represented their notion of the compound and 2) to rate the output from DALL-E in general. From the pictures generated by the AI system and the children, we can infer interesting differences in interpretation. Also, the automatically generated images only weakly represented the notion of the novel compound according to the participants. On average, the images generated were given a score of 2.4 (out of 5), where 1 means that the participant did not find the generated image to be representative at all, while 5 would mean a good representation. At the conference, we would like to present the dataset as well as some quantitative and qualitative analyses.

Keywords: Compounds, Compositionality, Image generation, Language processing

Baroni, M., Matiasek, J., & Trost, H. (2002). Predicting the components of German nominal compounds. In *ECAI* (Vol. 2002, p. 15th).

Dayma, B., & Cuenca, P. (2023). DALL·E Mini: Generate images from any text prompt. *Weights and Biases*. Retrieved February 23, 2023, https://wandb.ai/dalle-%20mini/dalle-mini/reports/DALL-E-mini-Generate-images-from-any-%20text-prompt--VmlldzoyMDE4NDAy

Shwartz, V., & Waterson, C. (2018). Olive oil is made of olives, baby oil is made for babies: Interpreting noun compounds using paraphrases in a neural model. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 2* (Short Papers), 218–224. https://aclanthology.org/N18-2035

9 Discourse Coherence in German and English Texts Produced by ChatGPT

Anastassia Shaitarova, University of Zurich

Martin Volk, University of Zurich

ChatGPT (OpenAI, 2022) is a state-of-the-art natural language processing model, which was released in November 2022 and belongs to the Generative Pretrained Transformer family (Brown et al., 2020). The model has received wide-spread attention due to its unprecedented performance, producing output that is almost indistinguishable from human-written text. Since such technology has the potential to affect the way we use language, there is a growing need for an extensive analysis of its capabilities as well as the development of a reliable method to distinguish the generated output from human text. Being part of the Language In The Human-Machine Era (LITHME) network, which investigates the influence of language technology on human language (Sayers, Sousa-Silva, & Hohn, 2021), we decided to run a fine-grained analysis of the ChatGPT capabilities. Inspired by the study of explicit connectives in language models (Beyer, Loáiciga, & Schlangen, 2021), we quantitatively investigated the usage of discourse particles in human-written and generated texts. Our hypothesis was that humans produce more coherent text, using more discourse particles than a pretrained model. We selected 50 texts from our multilingual banking magazine corpus (Volk, Amrhein, Aepli, Müller, & Strobel, 2016) in English and the corresponding ones in German (years 2002-2017). For every text we used the title and the first paragraph as a prompt for ChatGPT, prepending the command "Complete the text with about 500 words." ("Vervollständige den Text mit etwa 500 Wortern." for German texts) to each prompt, and clearly marking the title. For each text we truncated its human-written counterpart to match the length of the generated text. We then ran an extensive lexical analysis of the texts. In particular, we counted the number of discourse particles in each ChatGPT continuation and compared this to the number of discourse particles in the original texts in the same text span. We used the discourse connectives listed in the appendix of (Meyer, 2014). Our results show that the number of discourse particles in the human texts clearly exceeds the number in the ChatGPT-generated texts. For English, we counted a total of 822 (215 of them at the start of a sentence) discourse particles in the human texts and 653 (176) in the ChatGPT texts. For German, the ratio is 811 (192) in the human texts vs. 623 (159) in the automatically generated texts. Despite this clear tendency we observe that ChatGPT uses some connectives more frequently than humans (e.g. DE: trotz 16 vs. 6, and EN: in addition 21 vs. 8). We conclude that ChatGPT produces texts that are not as coherent as professionally written texts with respect to discourse particles. Thus, differences in the usage frequencies of discourse connectives can be used as an easily-accessible feature for the identification of ChatGPT-generated text, which is in line with the findings of (Dou, Forbes, Koncel-Kedziorski, Smith, & Choi, 2022). At the conference, we will present more detailed results, including other metrics. We offer to share our lists of discourse connectives and the 50 texts in English and German.

Keywords: ChatGPT, Discourse connectives, Identification of machine-generated text, Coherence

Beyer, A., Loáiciga, S., & Schlangen, D. (2021). Is incoherence surprising? Targeted evaluation of ´ coherence prediction from language models. *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (p. 4164–4173).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Dou, Y., Forbes, M., Koncel-Kedziorski, R., Smith, N. A., & Choi, Y. (2022, May). Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 7250–7274). Dublin, Ireland: *Association for Computational Linguistics*. Retrieved 2023, January 11, https://aclanthology.org/2022.acl-long.501

Meyer, T. (2014). Discourse-level features for statistical machine translation (PhD Thesis). *EPFL and Idiap Research Institute, Lausanne and Martigy*.

OpenAI. (2022, November). ChatGPT: Optimizing Language Models for Dialogue (Tech. Rep.). Retrieved 2023 February, 20, https://openai.com/blog/chatgpt/

Sayers, D., Sousa-Silva, R., Höhn, S., Ahmedi, L., Allkivi-Metsoja, K., Anastasiou, D., ... & Yayilgan, S. Y. (2021). The Dawn of the Human-Machine Era: A forecast of new and emerging language technologies.

Volk, M., Amrhein, C., Aepli, N., Müller, M., & Ströbel, P. (2016, September). Building a Parallel Corpus on the World's Oldest Banking Magazine. Bochum. Retrieved 2022, May 9, https://www.zora.uzh.ch/id/eprint/125746/.

Language rights

10

Stochastic Parrots, Storytelling and Behavioural Complexification: Throw Away After Use!

Faye Oosterhoff, Personal Interest

Development in speech technology is effectively interconnecting language information elements for text, spoken word and imagery. This is made possible because big data, stochastic algorithms, calculative speeds and its use for Natural Language Processing (NLP) have become part of the public domain. Connecting these worlds of information, seems advantageous as they originate from different context and modalities. It may consolidate different views for the same information. Besides, it would support those who have difficulty using either visual or aural sources for information. These so-called AI systems will rehash content keeping regenerated relational implications intact. In the same way, rendered AI images show very convincing 'novel' pictures based on database 'originals' containing descriptive information for content and style. As such, rather convincing AI storytellers are created to inform us. Storytelling also is our natural way to share memory. Whether it be Lasceaux cave images, psalms, pop music, family events, or this text, we all tell stories to remind and educate ourselves. In short, we're storytellers or 'Homo fabulans'. Computers are able to render story like communication without human intervention, but these systems have no real knowledge of their renderings interpretation. A passive 'bias' is created as a result of information recreated, based on conjecture. As such, content bearing on aggression, discrimination and human depreciation, will be put forward without taking into account semantic sensitivities. Companies jumping on this AI bandwagon, trying to meet public curiosity for the 'storytelling magic' involved, will see themselves manually deleting bias factors or add semantic metadata. The latter will reduce senselessness, but would also result in need for larger databases and calculation power. As for storytelling, there's still no actual dialogue, only a cleansed AI storyteller. (Bender and Gebru et al, 2021, p. 616) coined the term 'Stochastic Parrot'. A very apt description for AI storytelling. Humans share each other stories based on input they gather in their life, having similar quality. The question is, what makes us different from these language rendering systems? Well, we don't remember everything verbatim and we don't actually rehash on language. We formulate on understanding, interest, associations and feelings. As such, we add contextual information about what we think we individually represent at that time, relating to the situation we're in. Human dialogue is shaped as intersectional communication, differences meeting each other, staying different. They are building mutual rapport and trust, but don't average. AI storytelling represents a behavioural simulant, adding to human stories without actual human dialogue. Effectively, this devaluation of meaning is a bias too. It will make our behavioural world more complex by increasing misunderstanding, by cumulating apparent meaningful information sources. Thus, a behaviourally complexifying informational load created by AI contaminates human storytelling, as humans hardly recognise the difference. My suggestion would be: 'Throw away AI storytelling products after use.' Besides, AI storytelling also risks auto-contamination, using its own products, as put forward by (Bender and Gebru et al, 2021, p. 619).

Keywords: AI, Storytelling, Accountability, Semantics, Contamination

Bender E.M., Gebru T., McMillan-Major A., & Shmitchell, S. (2021, March 3-10). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? n *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

NLP

12 Constructing a framework for emergency medicine retrieval

Maria Onyshchuk

The report examines the emergency medicine terminology, and aims to describe a conceptual framework that presupposes the emergency medicine data retrieval by isolating a subset of texts from a larger corpus and basic text processing. This study presents my attempt to discuss both single and multi-word terms by means of statistical data. The material used for the research mostly contains the texts on military medical rescue operations with the focus on terminological units in English and Ukrainian. Methodologically, I apply the Sketch Engine tools to extract

terms, consider their frequency, analyse collocations within the occurrences. In my paper I argue that there are difficulties in the terminological units representation, which may be explained by the structural composition of the contrasted languages and the extensive elaboration of the English language in comparison with the Ukrainian. For this purpose, the results of the corresponding terminological lists, result retrieval and result ranking distributed word representations are considered. This study might help the researches in the domains of contrastive and applied linguistics, terminology of occupational languages, specialized discourses with a focus on military orientation of medicine, namely topic-related terminological units within the group of Tactical Combat Casualty Care (TCCC).

Keywords:

Anonymous, A. (2022). Tactical Combat Casualty Care (TCCC) Guidelines for Medical Personnel 15 December 2021. *Journal of special operations medicine: a peer reviewed journal for SOF medical professionals*, *22*(1), 11-17. https://pubmed.ncbi.nlm.nih.gov/35278312/

Biber, D., Conrad, S., & Reppen, R. (2011). *Corpus linguistics: Investigating language structure and use.* Cambridge University Press.

McEnery, T. (2001). *Corpus linguistics.* Edinburgh University Press.

Meyer, Ch. (2002). *English corpus linguistics. An introduction.* Cambridge University Press.

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge University Press.

Kunilovskaya, M. & Koviazina M. (2017). Sketch Engine: A toolbox for linguistic discovery. *Journal of Linguistics*, *68*(3), 503-507.

Thomas, J. (2016). Discovering English with Sketch Engine: a corpus-based approach to language exploration. Versatile.

Тактична медицина для підрозділів спеціального призначення. (2016). Київ: Начально-тренувальна база Атек. Рекомендації з тактичної допомоги пораненим в умовах бойових дій для медичного персоналу від 15 грудня 2021 року https://aaukr.org/rekomendatsiyi-z-taktychnoyi-dopomogy-poranenym-v-umovah-bojovyh-dij-dlya-medychnogo-personalu/

Scketch Engine text corpora analysis tool (https://www.sketchengine.eu/)

On the effectiveness of end-to-end text-to-speech for Persian language

Reihaneh Amooie

Text-to-Speech (TTS) systems, (also known as speech synthesis systems), are a fundamental component of human-machine interaction systems. To build a TTS system, various approaches have been proposed. The state-of-the-art methods use deep neural networks (DNNs) to implement end-to-end speech synthesis models using <audio, text> pairs in the absence of labeling information and a separate grapheme-to-phoneme model. In this paper, we implement and evaluate an end-to-end, DNN-based text-to-speech model using Tacotron (an end-to-end speech synthesis system by Google) for the Persian language to examine the effectiveness of this approach for this data-scarce language. In order to do so, we built a speech dataset of 49.8 hours by trimming Persian audiobooks into sentences. The audiobooks were narrated by a single professional female native Farsi speaker. The dataset comprises of 27999 short audio files and their transcripts. We used Mean Opinion Score (MOS) to evaluate the test samples based on their quality, intelligibility, and naturalness. For testing, 21 subjects listened to the samples and rated them 1 to 5. The scores for the models trained with a dataset of 25 hours and a dataset of 49.8 hours were 3.18 and 3.38, respectively. We compared the results from the experiment with a Persian TTS model (Ariana robot) based on statistical parametric methods trained on around 4 hours of data that utilizes a conventional pipeline architecture. In order to evaluate the model, we used different types of sentences, namely declarative, interrogative, imperative sentences, phrases and poems. The model performed differently on different sentence types, which can be attributed to the bias existing in the training dataset. Due to the unique complexities of the Persian language such as the absence of diacritics, presence of an abundance of homographs, and Ezafeh, end-to-end models trained on such amount of data do not seem to be an appropriate choice for Persian as a low-resource language. In the future works, we intend to examine transfer learning and data augmentation techniques in order to improve the quality of the output speech.

Since text-to-speech systems are an essential element in human-machine interaction systems, understanding the extent to which end-to-end deep neural networks can perform well in the presence of relatively small data can provide researchers, developers and computational linguists who are interested in the field of TTS with a clearer insight for choosing the best architecture for their models. Therefore, we think that the presented work relates to the aims and themes of LITHME.

Keywords:

Arık, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., ... & Shoeybi, M. (2017, July). Deep voice: Real-time neural text-to-speech. In *International Conference on Machine Learning* (pp. 195-204). PMLR.

Abutalebi, H. R., Bijankhan, M. (2000). Implementation of a text-to-speech system for Persian language. In *Sixth International Conference on Spoken Language Processing.*

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Behbahani, Y. M., Babaali, B., Turdalyuly, M. (2016). Persian sentences to phoneme sequences conversion based on recurrent neural networks. *Open Computer Science, 6*(1), 219-225.

Bińkowski, M., Donahue, J., Dieleman, S., Clark, A., Elsen, E., Casagrande, N., ... & Simonyan, K. (2019). High fidelity speech synthesis with adversarial networks. *arXiv preprint arXiv:1909.11646*.

Black, A. W. (2006) CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling. In *Ninth International Conference on Spoken Language Processing*.

Daniel, J., James, H. M. (2009). *Speech and Language Processing*. Prentice Hall.

Farrokhi, A., Ghaemmaghami, S., Sheikhan, M. (2004). Estimation of prosodic information for Persian text-to-speech system using a recurrent neural network. *In Speech Prosody 2004*, *International Conference*.

Griffin, D.; Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process. 32*, 236–243.

Homayounpour, M. M., and Namnabat, M. (2007). FARSBAYAN: A Persian speech synthesizer based on unit selection method. *The CSI journal on computer science and engineering*.

Hunt, A. J., & Black, A. W. (1996, May). Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings* (Vol. 1, pp. 373-376). IEEE.

Javidan, R., & Rasekh, I. (2010). Concatenative Synthesis of Persian Language Based on Word, Diphone and Triphone Databases. *Modern Applied Science*, *4*(10), 97.

Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, *27*(6), 349-353.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008, March). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *2008 IEEE International Conference on Acoustics, Speech and Signal processing* (pp. 3933-3936).

Koriyama, T., Nose, T., Kobayashi, T. (2013). Statistical parametric speech synthesis based on Gaussian process regression. *IEEE journal of selected topics in Signal Processing, 8*(2), 173-183.

Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M. (2019, July). Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence 33*(01), 6706-6713.

Morise, M. (2015). CheapTrick, a spectral envelope estimator for high-quality speech synthesis. *Speech Communication*, *67*, 1-7.

Naderi, N., Nasersharif, B., Nikoofard, A. (2022). Persian speech synthesis using enhanced tacotron based on multi-resolution convolution layers and a convex optimization method. *Multimedia Tools and Applications*, 1-17.

Naiemi, F., Ghods, V. (2019). Persian Speech Synthesis Using Pitch Frequency in Flite software. *Advanced Signal Processing, 3*(1), 97-107.

Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T. Y. (2019). Fastspeech: Fast, robust and controllable text to speech. *arXiv preprint arXiv:1905.09263*.

Shannon, M., Zen, H., Byrne, W. (2012). Autoregressive models for statistical parametric speech synthesis. *IEEE transactions on audio, speech, and language processing, 21*(3), 587-597.

Shen, J., Pang, R. M., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z. H., ... & Wang, Y. X. (2018, April). RS-Ryan,". *Natural tts synthesis by conditioning Wavenet on mel spectrogram predictions*. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4779-4783). IEEE.

Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A., Bengio, Y. (2017). Char2wav: End-to-end speech synthesis. In *ICLR2017 workshop submission*.

Vainer, J., Dušek, O. (2020). Speedyspeech: Efficient neural speech synthesis. *arXiv preprint arXiv:2008.03802*.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Yin, X., Ling, Z. H., Dai, L. R. (2014, May). Spectral modeling using neural autoregressive distribution estimators for statistical parametric speech synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3824-3828). IEEE.

13

Machine-led concept extraction

Justyna Robinson, University of Sussex

Sandra Young, University of Sussex; Rhys Sandow, University of Sussex; William Kearney, University of Sussex

At Concept Analytics Lab we have developed a novel approach to text analysis which allows the machine to read large amounts of texts and extract the keyconcepts from that text. This approach also visualizes key concepts through a bespoke web application and a visualization tool. The approach extracts conceptual information from a text using WordNet (Princeton University 2010). WordNet puts a taxonomic understanding over a text using lexical inheritance system (Miller 1995, Fellbaum 1998). In order to identify keyconcepts we use Pointwise Mutual Information (PMI). PMI allows to determine which concepts are distinctive of the target corpus with the reference corpus. Exploring these concepts through our visualisation tool enables us to discover conceptual patterns which are otherwise hidden when traditional text analysis techniques are employed. These patterns allow researchers to meaningfully direct further analysis which includes more traditional corpus or narrative techniques. In some respects this approach shares affinity with other information extraction computational approaches such as topic analysis. However, the principles on which this approach is build differ from existing approaches. Because our approach aggregates words into concepts according to their shared meaning, individual words, which in topic analysis would not reach the statistical threshold, are picked up by our approach when they are a part of a statistically significant group of words. By broadening the focus from a word to concept we are able to demonstrate the salience of the entire semantically-related group of words that comprise a given concept. This approach allows extracting the topics of discourses as opposed to the form of the words that are used to express those topics. In our presentation we showcase this approach by extracting and analysing salient concepts from Covid-era and use previous nine years of data as a reference corpus. The data comes from an archive of day diaries collected as part of the Mass Observation (http://www.massobs.org.uk/) written on the 12th of May on each of the years from 2010 to 2020. We conclude by assessing the approach and outlining next set of challenges in machine-led concept extraction research.

Keywords: Concept, Variation, Covid

Concept Analytics Lab (https://conceptanalytics.org.uk/)

Mass Observation Project, The. 1981–Present. (http://www.massobs.org.uk/). Accessed 2023, February, 2.

Miller, Miller, G. A. (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

Fellbaum, Christiane (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Princeton University. 2010. About WordNet. (https://wordnet.princeton.edu/) Accessed 2023, February, 2.

14

The Impact of Speaking Style on Fricative Detection: A Machine Learning Approach

Nina Hosseini-Kivanani, University of Luxembourg

Homa Asadi, University of Isfahan; Christoph Schommer, University of Luxembourg

The unique characteristics of speakers' voices make it possible to identify them based on their voice patterns. Speakers' voices are highly individualized, and frequently vary more within the same speaker than between speakers. Because of this variability, it is challenging for both humans and machines to differentiate between speakers (Hansen, J. H., & Hasan, T., 2015). Cho et al. (2020) conducted a cross-linguistic study on the perception of fricatives in clear speech, comparing native speakers of English and Korean. They highlighted the significance of considering language-specific factors and the effect of clear-speech on multi-modal perception of English sibilant and non-sibilant fricatives. Maniwa (2007) and Maniwa et al., (2009) examined the acoustic characteristics and perception of clear and conversational speech modes of fricatives in English. They investigated the variations in the spectral and temporal properties of the fricatives using acoustic analysis. They found that the speaking style has a significant impact on the acoustic characteristics of the fricatives. To address this challenge, we have performed an evaluation study where different machine learning models have been developed to identify the best classification model, i.e., Naïve Bayes (NB), Support Vector Machines (SVM), Random Forest (RF), and K-nearest neighbors (KNN). These models were implemented using the scikit-learn library (Pedregosa et al., 2011) in Python (3.9). The study aimed to detect fricatives, in four speaking style datasets (i.e., read, clear, spontaneous, and child speech) by considering acoustic features of center of gravity, standard deviation, and skewness. The recorded dataset consists of speech samples from 40 male Persian speakers, who spoke in four different speech styles and were balanced in age. As the number of speech samples for each segment was not identical in our dataset, a balancing technique was used to create a dataset with the same number of samples for each segment of fricatives. The imbalance-learn library from the contrib packages of scikit-learn was used to balance the dataset. After preprocessing the data, k-fold cross validation was used to optimize the model's output and splitting it into training and testing sets. To enhance the model's performance, several ML models were trained. The findings showed that RF model for clear and read speaking styles is the best performing model for predicting fricatives in different speaking styles, with accuracy of 73,3% for the clear speaking style and 72% for the read speaking style. Our results also showed that the voiceless fricatives /f/, /s/, and /ʃ/ were best classified by the RF model, whereas the voiced fricatives /v/, /z/, and /ʒ/ and voiceless fricative /h/ and /x/ were poorly classified by the model. This may indicate that the voiceless fricatives /f/, /s/, and /ʃ/ are more speaker-specific than others according to this data. These findings have important implications for speech therapy and communication disorders and are relevant to LITHME's mission to explore new applications of machine learning and artificial intelligence in various fields.

Keywords: Machine learning, Acoustic features, Speaking style, Detection

15 Machine helping IsiXhosa speakers read handwritten notes

Justyna Robinson, Champion Mobile Global

Yassir Laaouach, Champion Mobile Global; Gayathri Sooraj, Champion Mobile Global

Champion Mobile Global Ltd is a UK company focused on delivering software solutions to barriers in human communication. We focus on the linguistic situation in South Africa which juggles 11 official languages and relatively low literary problems. In this presentation we focus on one of the challenges we have been working on, namely machine reading of a handwritten text. Out team has made significant strides in advancing the field of OCR technology. Specifically, we have developed an innovative OCR system that is specifically tailored to one of the South African languages, i.e., IsiXhosa. Additionally, we have incorporated advanced language models into our OCR pipeline in order to enhance the semantic understanding of our OCR system, resulting in a more accurate and reliable output. The initial phase of development focused on the establishment of the core OCR pipeline, including the implementation of state-of-the-art techniques for text recognition. In order to accurately recognize handwritten and typed text, our team has incorporated an AI automated recognition tool into our OCR technology. Handwritten text, in particular, presents a unique challenge due to the rich variance in styles. The subsequent phase was primarily dedicated to the collection and annotation of a large dataset of training data for IsiXhosa, which was crucial for the fine-tuning and optimization of our OCR system. One of the main challenges faced in this project was the lack of resources available for low-resource languages such as IsiXhosa. Traditional OCR systems often struggle to accurately recognize text in these languages due to a lack of training data and poor image quality. By incorporating context-aware correction into our OCR pipeline, we were able to overcome these challenges and achieve a high level of accuracy in recognizing text in IsiXhosa. Moreover, IsiXhosa have a more complex grammatical structure, which includes noun classes, concords, and complex verb conjugations. Due to the complex grammatical structures of Bantu languages like IsiXhosa, traditional OCR systems may struggle to accurately recognize text in these languages. The lack of resources available for these languages and the poor image quality often make it difficult for the OCR systems to understand the context of the text, which in turn leads to errors in the recognition process. In summary, our team has developed an innovative OCR solution that addresses the specific needs of low-resource languages, while also incorporating advanced techniques such as context-aware correction and language modeling to enhance the accuracy and reliability of our OCR system. By using English as a test case, we were able to optimize and test our contextual OCR model before applying it on the selected languages. This technology has the potential to greatly benefit communities that rely on these languages and have limited access to existing OCR solutions.

Keywords: ICR, IsiXhosa, Language model, South Africa

Interpretation

16

Interpreting awareness in Estonian Public Service Interpreting

Jekaterina Maadla, Tallinn University

While having only one official language, the linguistic landscape of the Republic of Estonia is diverse and historically complex. Estonian is spoken as a first language by 63% of the population (Statistics Estonia, 2021), and, although the proficiency in Estonian among youngsters with a different first language is high 71% (Lukk et al., 2017, p. 14), Estonia remains linguistically divided and mutual understanding in different regions of the country can be limited, resulting in a steadily growing role for non-Estonian languages in the institutional framework and public administration services. Failure to pay attention to this fact and its implications, however, can lead to an array of problems that put quality of public services at risk (Council of Europe, 2022, p. 16), thus jeopardising communication with linguistic minorities during the COVID pandemic and amidst Ukrainian war crisis. One of the main challenges in this respect is that commissioners in public services lack awareness of the role and implementation of interpreting. Because of the widespread of multilingualism in Estonia, where one in two inhabitants speaks a foreign language — 49% English and 39% Russian (Statistics Estonia, 2021), interaction between state authorities and non-Estonian service users is often mediated by ad hoc interpreters and multilingual staff. This has led to a critical shortage of professional interpreters and poor interpreting policies, with Public Service Interpreting (PSI) being one of the least developed fields in Estonia when it comes to interlingual communication. This study aims to explore the role of language mediation in Estonian public services, how Estonian public service commissioners understand language mediation or interpretation, its use and implications, and what obstacles we might encounter when raising awareness of the need for quality interpretation to ensure a viable, ethical and fair use of professional, ad hoc and volunteer interpreters. Although Estonia is currently developing AI-based tool for multilingual public communication Bürokratt (RIK, 2023) and the project focused on creating a central translation motor is on the run (Wright, 2021), awareness of the role and implementation of interpreting within commissioners of these tools is low. Additionally, the bias against CAT-tools within interpreters and translators is also a factor limiting development of viable solutions for CAI in the public sector. To overcome these challenges, more research is needed to evaluate opportunities and challenges of implementation of CAI and CAT-tools in public service. For this purpose, and as a first step in mapping the field of under-researched topic of PSI in Estonia, I conducted qualitative interviews (Magnusson & Marecek, 2015) with heads of communication of local and national public service institutions and non-governmental agencies involved in provision of interpreting services and multilingual communication. This study will serve as a foundation for a broader one delving deeper into the actual interpretation practices by professional and ad hoc interpreters and awareness about using machine translation tools for linguistic mediation.

Keywords: Linguistic inclusion, Metalinguistic awareness, Public service interpreting

Council of Europe. (2022). Advisory committee on the framework convention for the protection of national minorities. Secretariat of the Framework Convention for the Protection of National Minorities. https://rm.coe.int/5th-op-estonia-en/1680a6cc9e

Lukk, M., Koreinik, K., Kaldur, K., Vihman, V.-A., Villenthal, A., Kivistik, K., Jaigma, M., & Pertšjonok, A. (2017). Eesti keeleseisund. Tartu Ülikool (RAKE), Balti Uuringute Instituut (IBS). http://hdl.handle.net/10062/56511

Magnusson, E., & Marecek, J. (2015). Doing interview-based qualitative research: *A learner's guide*. Cambridge University Press.

Määttä, S. (2015). Interpreting the discourse of reporting in interviews conducted by law enforcement agents. *The International Journal of Translation and Interpreting Research*, *7*(3).

RIK (Republic of Estonia Information System Authority). (2023). Bürokratt. https://www.ria.ee/en/state-information-system/machine-learning-and-language-technology-solutions/burokratt.

Accessed 15.01.2023

Statistics Estonia. (2021). Demographic and ethno-cultural characteristics of the population. https://rahvaloendus.ee/en/results/demographic-and-ethno-cultural-characteristics-of-the-population

Wright, H. (2021, July 9). Estonia to create a central translation platform. Estonian Broadcasting Corporation ERR. Accessed 15.01.2023. https://news.err.ee/1608272391/estonia-to-create-a-central-translation-platform

17

When you have "mastered" numbers: Language interpreters in the human-machine era -- a case study

Jun Pan, Hong Kong Baptist University

Yucheng Jin, Hong Kong Baptist University

Simultaneous interpreting, wherein listening and speaking of two languages occur at the same time, is considered a high cognitive demand task. Research has proven that numbers constitute one of the greatest contributors to increased cognitive load for interpreters. The problem is getting more prominent when interpreters work between languages with entirely different numeric systems such as Chinese and English. In recent years, computer-assisted interpreting (CAI) tools have been developed (e.g. InterpretBank) to facilitate interpreters by providing automatically processed information (e.g. transcription, extraction of numbers and proper names, and term translation). Nevertheless, the application of CAI tools during interpreting has led to mixed feedback. The introduction of extra visual input by CAI tools, when not well contextualised, may lead to increased cognitive loads. Among the various features of CAI tools, automatic speech recognition (ASR) generated numbers have in general been regarded as useful and constitute an important feature language interpreters in the human-machine era may benefit from. Still, the potentially increased cognitive load created by the extra visual input during highly intensive activities such as simultaneous interpreting should be carefully looked into. This study will, therefore, from an ergonomics point of view, investigate the design of AI systems to facilitate number rendition for language interpreters in the human-machine era, in the hope of shedding light on enhanced human-machine interaction at large. In particular, the study set out to investigate how an AI-powered system could help interpreters recognise and memorise long-digit numbers. According to the outputs of ASR and Machine Translation technology (MTT), we propose three experimental conditions that provide AI assistance differently. A total of 12 student interpreters will be randomly assigned to simultaneously interpret texts carefully designed with large numbers. Specifically, a control group of 4 will interpret numbers without getting any AI assistance from the system. The other two experimental groups (4 each) will receive ASR assistance and ASR+MTT assistance, respectively. For example, given a task of translating 2,333,123 from English to Chinese, the output of ASR assistance is 2 million 333 thousand and 123; and ASR+MTT assistance presents the number in the Chinese digit format that is 233 wan 3qian 1bai 23. Based on the three experimental conditions, we design a within-subject study to compare which system assistance is most favorable regarding effectiveness and cognitive load. We will use a mixed method combining quantitative and qualitative data to analyse user behaviour and their perceptions of the AI assistance. We will use eye-tracking data (e.g. fixation duration, eye movement patterns) to capture how interpreters actually treated the AI assistance, and adopt a self-reported questionnaire (NASA-TLX) to analyse their cognitive load when incorporating the AI assistance into the interpretation task. Results of the study can help to unveil the "black box" of human interpreters in complex and cognitively demanding tasks such as simultaneous interpreting. The study will also provide insights to the ergonomic design of AI tools that can facilitate human performance in high-level thinking tasks.

Keywords: Simultaneous interpreting, Cognitive demand, Human-machine interaction, Computer-assisted interpreting, Numbers

This study is sponsored by HKSAR's Research Grants Council under its General Research Fund (Project Number: 12623122).

Translation

18

Teaching legal translation at MA level with the assistance of neural machine translation

Réka Eszenyi, senior lecturer, Eötvös University, Budapest

The objective of the action research project is to investigate the processes and products of an experimental translation seminar where students translated, post-edited and revised legal texts from Dutch into Hungarian. The translator trainees took turns in playing the role of the translator, the post-editor of the output of a neural machine translation (NMT) engine, and revisor of the two types of translation in 10 different assignments during a semester. Is it worth the effort to use NMT for these texts in this language combination? What are the advantages and drawbacks of translating in the two different modes: relying on human and machine translation? When it comes to post-editing, is it worth to post-edit machine-translated texts, or can the raw output be used, and when it comes to revision, to revise post-edited texts? How has the role of the translator and her array of competences changed? The study focuses on these questions by investigating the translations produced for the classroom assignments in the genre of legal texts, in the Dutch-Hungarian language combination, a segment of the translation market that may seem tiny, yet it is very hard to find qualified translators who take such assignments. Can NMT be employed to ease the workload of legal translators? If yes, what sort of changes are needed in translator training, and what competences do the translators of the (near) future need? As of the year 2016, the most widely used form of automated translation is neural machine translation a form of machine translation that uses a large artificial neural network to make predictions about the probability of a sequence of words. Using NMT, it is possible to produce much more precise and natural-sounding texts compared to earlier rule-based or statistical systems. These earlier systems could also be helpful in translating certain genres between languages that are quite close to each other. But for medium-sized languages like Dutch, and relatively isolated languages like Hungarian, which among the official languages in Europe is only related to Finnish, and there is no mutual intelligibility there either, the emergence of NMT marked a new chapter in translation. These changes brought about an increase in post-editing activity (ELIS 2022). The task of the post-editor is similar to that of the revisor from several perspectives (Daems & Macken, 2020), although the level of post-editing can vary greatly depending on the desired quality of the translation.

Keywords: Neural machine translation, Translation training, Post-editing, Revision, Legal translation

Daems, J., & Macken, L. (2020). Post-editing human translations and revising machine translations: impact on efficiency and quality. In *Translation Revision and Post-Editing* (pp. 50-70). Routledge.

ELIS (EUROPEAN LANGUAGE INDUSTRY SURVEY) (2022). https://ec.europa.eu/info/sites/default/files/about_the_european_commission/service_standards_and_principles/documents/elis2022-report.pdf.

19

CAT tools in revoicing - Are they a help or a hindrance in dubbing and voice-over translation?

Márta Juhász-Koch, Eötvös Loránd University

In 2021 Audiovisual Translators Europe issued a Machine Translation Manifesto (Deryagin, Pošta and Landes, 2021), urging the language industry to implement augmented translation tools into the field of audiovisual translation (AVT). But would they be useful in all the major modes of AVT? Although subtitling has already been somewhat facilitated by the use of translation memory components used in could-based platforms such as GTT and Transifex as well as in translation softwares such as MemoQ, SDL Trados Studio and Transit NXT (Athanasiadi, 2015), very few considerations about revoicing (dubbing and voice-over translation) have been taken into account in CAT tool development. A likely reason for this is because unlike subtitling, dubbing and voice-over scripts are still not translated in specified software platforms but most often in unique Word document templates issued by the dubbing studio. Among other aspects of machine translation the manifesto advises that more traditional CAT tools ought to be customized for AVT, notably translation memory (TM), termbase (TB) and voice input (VI). The use of translation memories would undeniably be beneficial in both the dubbing of fiction and the voice-over translation of non-fiction programmes, especially if the product is a series of episodes. Dubbing and voice-over translation would also benefit from consistently accessible termbases, i.e. when accepting an assignment which is part of a bigger project such as a documentary or a fiction series, the translator would automatically be provided with a termbase. Consistency Sheets are sometimes used by dubbing studios in a shareable and editable Excel format but more often than not the translator is merely given a translated script of an earlier episode of the series to peruse (which is not only time consuming but one translation is an insufficient source). A systematic, searchable term base complemented by predictive typing would definitely enhance the speed, efficiency and accuracy of the revoicing process. But the most beneficial CAT tool indicated in the manifesto seems to be the voice input function. In both dubbing and voice-over, before submission the translator is required to review the whole target language script as if they were the voice actor at the production's actual voice recording in order to test the length of the target-language segments. This is an additional, sometimes time-consuming phase of the translation process often requiring heavy post-editing, and it also requires routine from the translator in managing to imitate the voice actor's speed and proficiency. A voice input or dictation function as a CAT tool would incorporate the read-aloud phase into the actual translation, speeding up the whole process. However, all three CAT tools raise questions about certain problematic aspects typical of revoicing, e.g. the time constraints of certain segments not allowing the use of solutions provided by TMs, or the accuracy and maximum speed of VI which has to keep up with the actors' tempo. In my presentation I am going to investigate the advantages and inconveniences of these CAT tools from the perspective of the language professionals' work in revoicing.

Keywords: Audiovisual translation, Revoicing, Translation memory, Termbase, Voice input, Dubbing, Voice-over translation

Athanasiadi, R. (2015) Applications Of Machine Translation And Translation Memory Tools In Audiovisual Translation: A New Era? KantanMT News website: https://kantanmtblog.com/2015/10/05/applications-of-machine-translation-and-translation-memory-tools-in-audiovisual-translation-a-new-era/ Retrieved: 20 Feb 2023.

Deryagin, M., Pošta, M. and Landes, D. (2021) AVTE Machine Translation Manifesto. Audiovisual Translators Europe website: https://avteurope.eu/wp-content/uploads/2022/10/Machine-Translation-Manifesto_ENG.pdf Retrieved: 20 Feb 2023.

20 The changing role of subcompetencies in the translator's work

Márta Lesznyák, University of Szeged

It is widely accepted these days that translation as a profession is undergoing profound change. It is clear that in the everyday work of translators, post-editing is gaining ground and pushing traditional human translation back, even if context, language pair, text-type and other factors may have an impact on how a specific translation task is carried out. It is also clear that new work modes require new skills and may change the relative weight of subcompetencies that were once determined as the cornerstones of translation competence. Although there is a lot of speculation on what skills and subcompetencies translators will need in the Human Machine Era (e.g. Rico and Torrejón, 2012; Pym, 2013, Nitzke et al. 2019), there is a dearth of empirical research on the topic, particularly on basic subcompetencies seemingly not directly involved in specific post-editing tasks. The Translation and MT Post-Editing Competence Research Group of the University of Szeged, Hungary started a research project in 2020 with the aim of discovering whether traditional translation sub-competencies as defined by the PACTE group (2003) play similar roles in post-editing as in human translation. Since 2020 six waves of data collection have been carried out, and data has been collected from 99 MA students of translation at the University of Szeged. The project includes and controls the following variables: source language skills (reading and grammar competence), thematic knowledge, beliefs about translation, text-type, translator's experience, method of translation (human translation /HT/ and post-editing /PE/) and students' perceptions of the advantages and disadvantages of working with HT and PE. In this presentation, the results of the legal translation condition (text-type controlled) will be summarized. Altogether 49 second-year MA students at the end of their studies participated in this part of the investigation. About half of the students translated and the other half post-edited a part of a copyright agreement (350 words) from English into Hungarian. The quality of the target texts was evaluated by three raters using an MQM-based error typology. In addition, the respondents filled in reading and use of English tests, a copyright test (to assess legal background knowledge), a questionnaire on beliefs about translation and a follow-up questionnaire right after submitting the target text on their perceptions of HT and PE. The results of the statistical analyses indicate that reading and grammar competence show significant correlations with both translation and post-editing performance. Moreover, reading's relation to performance in the post-editing condition is even stronger than in the human translation condition, suggesting that the role of reading is even more important in post-editing than in traditional human translation. Beliefs about translation showed significant correlation with several indices of human translation performance but only with one aspect of post-editing. No significant correlations were found between thematic knowledge and translation or post-editing performance. In addition, students' perceptions seem to contradict some of the empirical findings, suggesting that their subjective perceptions are not in line with their objective performance. Explanations will be provided for all the research findings listed above and we will argue that the importance of basic background competencies is shifting, which calls for a revision of what and how translation students should be trained. The research presented in this paper is closely related to the following themes and aims of WG7 LITHME: Training language professionals (translators), the (changing) role of humans in translation, machine translation of legal texts, language technologies in translation.

Keywords: Translation competence, Post-editing competence, Reading in post-editing, Thematic knowledge in post-editing, Beliefs about translation

References Nitzke, J., Hansen-Schirra, S., & Canfora, C. (2019). Risk management and post-editing competence. *The Journal of Specialised Translation 31*, 239–250.

PACTE. (2003). Building a Translation Competence Model. In F. Alves (ed) *Triangulating Translation: Perspectives in Process Oriented Research*, (pp. 43–66.) *Amsterdam: John Benjamins*.

Pym, A. (2013). Translation Skill-Sets in a Machine-Translation Age. *Meta*, *58*(3), 487–503.

Rico, C. & Torrejón, E. (2012). Skills and profile of the new role of the translator as MT post-editor. *Revista Tradumática 10*, 0166–178.

Vitality

21

Questioning the revitalisation of an endangered language in the Human-Machine Era The case of Valoc' in the Rhaetic Alps

Fabio Scetti, Université Paul-Valéry Montpellier 3

This contribution provides important insight into the complex issue of promoting the use of an endangered languages in the Human-Machine Era, when the Internet plays a major role in terms of visibility as well as being an easy and speed way of communication. This research focuses on the webpage Vocabolär del Valoc' de la Val Mäśen (VVV) which is part of the homonym project created in 2017. The aim of this project is to study language practices and representations of Valoc', a dialect of Lombard spoken in Val Masino, a lateral valley located in lower Valtellina (Northern Italy), and promote its use. The valley is populated by around 1,000 inhabitants, however Valoc' dialect is spoken only by older generation and by some other people who left Italy and live abroad, mainly in Argentina, the USA, Australia, France and Switzerland. The webpage represents then the link between these people connecting families that have been separated for decades. Our approach is both on dialectology and sociolinguistic as we complete our study with observations and interviews among speakers of different ages, sexes and professions, to see how Valoc' is still used and from whom. To conclude, this contribution allows us to reflect on how the new 'global' society may influence the process of transmission of this endangered dialect, which needs to be revitalized. The webpage was important in order to introduce Valoc' as a vehicular language not only orally but also in written form. Finally, we address the importance of developing a dictionary in order to promote a unique norm of reference, as a way of preserving Valoc' for the future.

Keywords: Valoc', endangered Language, Revitalisation, Language documentation, Human-Machine Era

22

Dialects in the human-machine era – between vitality and endangerment: Implications of crowdsourced language data generation

Barbara Heinisch, University of Vienna

Linguistic diversity is a major element of human cultures. While language variation is increasingly considered in language technology development, dialects are still under-represented. This is despite the fact that dialect loss (Schilling-Estes & Wolfram, 1999) is as prevalent as the loss of languages all over the world. Measures to counter this loss, preserve dialects for future generations or even promote the usage of dialects in the present are, among others, crowdsourcing initiatives aimed at the creation of language data for the further development of language technologies. Illustrated by a citizen science project in the field of lexicography (Heinisch, 2020), this presentation demonstrates how language data can be generated together with dialect speakers and further processed to make them available for the use by language technologies. The presentation assesses the underlying ethical, technological and societal implications of crowdsourced language data generation for further use in language technologies. Among the ethical implications is the instrumentalisation of dialect speakers that might be outweighed by the benefits they gain. Furthermore, there are challenges resulting from the crowdsourced preservation of cultural heritage. Moreover, the reversed role between researchers and laypersons regarding expertise in the case of dialects emphasises the significance of lived experience in dialect use. Also, the intended further usage of language data, such as for language technology development, give rise to ethical considerations. Among the technological implications are usability considerations. Since the persons creating language data in the case of citizen science are usually not experts in linguistics it is necessary to strike a balance between simplifying the user interface while guaranteeing high data quality. Additionally, further use of the created language data by humans and machines needs to be taken into consideration. This may include making the data available in a visually appealing way for the contributors themselves and making them FAIR (findable, accessible, inter-operable, re-usable), such as through the Linguistic Linked Open Data Cloud (Cimiano, Chiarcos, Mccrae, & Gracia, 2020), thus opening them up for a wide range of potential (language technology) applications. Societal implications of the crowdsourced creation of language data are the public perception of the endeavour itself, the effects on language technology development in everyday life and the impact on policy. The public perception of dialects, including ideologies and attitudes about language, has an effect on dialect vitality (Schneider, 2018). Therefore, the participation in and media coverage of dialect data initiatives may also influence the preparation of language policies and thus, stimulate the development of language technologies focusing on dialects.

Keywords: Citizen science, Language vitality, Language data, Language technologies

Cimiano, P., Chiarcos, C., Mccrae, J. P., & Gracia, J. (2020). Linguistic Linked Open Data Cloud. In *Cimiano & Gerstner (Eds.), Linguistic Linked Data* (1st ed., pp. 29–41). *Springer International Publishing.* https://doi.org/10.1007/978-3-030-30225-2_3

Heinisch, B. (2020). Developing Language Resources with Citizen Linguistics in Austria – A Case Study. In J. *Fiumara, C. Cieri, M. Liberman, & C. Callison-Burch (Eds.), Citizen Linguistics in Language Resource Development (CLLRD 2020) Proceedings.*

LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020 (pp. 7–14). European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/2020.cllrd-1.2/

Schilling-Estes, N., & Wolfram, W. (1999). Alternative Models of Dialect Death: Dissipation vs. Concentration. *Language*, *75*(3), 486–521. https://doi.org/10.2307/417058

Schneider, C. (2018). Ideologies, practices, and their effect on dialect vitality: a case study from Vanuatu. *Journal of Multilingual and Multicultural Development*, *39*(1), 60–75. https://doi.org/10.1080/01434632.2017.1311904

Low-resource/ minority languages

23 Cross-Lingual Transfer Learning for Low-Resource Languages

Fred Philippy, **Zortify S.A**.

It is widely acknowledged that Transformers have had a profound impact on the field of Natural Language Processing (NLP), and that their success has led to the development of a variety of large pretrained language models like BERT, RoBERTa, and T5. These models have achieved state-of-the-art performance on a variety of NLP tasks, but their development has primarily focused on the English language. In order to address this limitation, NLP researchers have started to build on the success of these monolingual models and to develop their multilingual counterparts. One such model is mBERT, which is a version of BERT that can process text in multiple languages. By leveraging data from multiple languages, pretrained multilingual language models (MLLMs) like mBERT have shown promising results in a variety of multilingual NLP tasks. Nevertheless, the imbalanced availability of text data across languages often results in a lack of representation for languages with limited resources in the pre-training corpora of MLLMs. As a result, MLLMs tend to exhibit poorer performance when processing text in low-resource languages. Although MLLMs like mBERT are pre-trained without the use of parallel corpora or any cross-lingual signal, they display remarkable cross-lingual transfer capabilities, providing an opportunity to mitigate the performance gap among languages. A common approach to harness the cross-lingual transfer ability of MLLMs is to leverage knowledge learned by the model in a high-resource source language (such as English) and to transfer it to a low-resource target language. This presentation will explore the inner workings of cross-lingual transfer and discuss different factors that impact cross-lingual transfer in MLLMs. Furthermore, it will provide a comprehensive overview of cross-lingual transfer in MLLMs and offer insights into how to effectively leverage this capability for low-resource languages. It attempts to present a fair and accurate representation of the current state of research on this topic, with the aim of initiating discussions and potentially facilitating the inclusion of novel perspectives of researchers from various language technology fields into the current research line. In accordance with one of LITHME's objectives of tackling the inequality that minority languages are confronted with, this presentation aims to contribute to the ongoing debate about the factors impacting cross-lingual transfer performance, with the ultimate goal of ensuring that low-resource languages can equally benefit from the use of MLLMs. It is important for all languages and communities to have equal access to the benefits and opportunities provided by the advances in natural language processing, and this presentation aims to serve as a useful resource in this regard.

Keywords: **Low-resource**, **NLP**, **Cross-lingual**, **Language models**

24

"Quo Vadis, regulators?" Media in minority languages meet AI

Tarlach McGonagle, Leiden Law School, Media Law and Information Society

Tom Moring, University of Helsinki, Swedish School of Social Science

This paper sets out to explore a selection of new and far-reaching questions in the light of opportunities and threats to minority languages in the human-machine era. It will focus in particular on regulation and policy relating to the use of minority languages in the media. The paper will draw on the authors' long-standing experience of academic and expert advisory engagement with relevant issues. As has been repeatedly shown (McGonagle and Moring, forthcoming 2023, the OSCE HCNM's Tallinn Guidelines 2019), regulation has all but followed the technological developments in the media field, leaving minority languages far behind a level of support that equipped them in earlier media generations. Optimists may point to opportunities for certain more resourceful languages and narrow groups of activists, with a take-up of computational linguistics and front-line use of AI. On the other hand, smaller language environments, lacking necessary resources in the form of digitalized data banks available for data mining, and a market big enough to support investments, lag behind. The human machine-era affects all aspects of society, which means that measures to avoid negative developments and seek to induce a positive turn must be based on a holistic approach, leaning on various innovative strategies. These would include (at least) legal and regulatory measures (further developing measures recommended under the ECRML (1992, and subsequent opinions), various OSCE media guidelines, etc.), policy measures that critically look at the core ingredients, "opportunity", "capacity" and "desire", as suggested by Grin and Villancourt (1999), and take a new approach to the very concept of "language" as a boundary-driven concept. This evidently also includes a new understanding of power-broking in, and between, language communities. In the past, the media - and social and other new media actors – have always contributed to the development and consolidation of emerging information and communications technologies. Regulation of the developing media often struggled to keep pace with technological change; old regulation was often repurposed for new technologies. This regulatory strategy of "adaptive replication" (McGonagle 2020) worked quite well for a long period, but the changes induced by the Internet of things and the machine-human era are so sudden and so forceful that we must urgently enquire whether a complete regulatory rethink is now necessary. In our effort to contribute to a reshaping of the media-policy map in the human-machine era, we wish to reach out to scholars working with language technology as well as to scholars working with legal measures and scholars working with attitudes and social psychology. We have seen it coming, but perhaps without appreciating its full force, speed and sophistication.

Keywords: Minority language and AI, Media regulation, Media policy, Language interaction

Council of Europe (1992). European Charter for Regional or Minority Languages (ECRML), CETS No. 148, 1992.

Grin, F. & Vaillancourt, F. (1999). The cost-effectiveness evaluation of minority language policies: Case studies on Wales, Ireland ant the Basque Country. *Flensburg: European Centre for Minority Issues*.

McGonagle, T., & Moring, T. (forthcoming 2023). Language Policy and Regulation in the Old and New Media. In *François Grin, Michele Gazzola, Linda Cardinal and Kathleen Heugh (Eds.), The Routledge Handbook of Language Policy and Planning.* Routledge.

McGonagle, T. (2020). Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation. In *Giancarlo Frosio (Ed.), The Oxford Handbook of Online Intermediary Liability.* Oxford University Press, pp. 467-485.

OSCE (2019). Tallinn Guidelines on National Minorities and the Media in the Digital Age. OSCE High Commissioner on National Minorities (HCNM).

25

Connecting the dots, efforts on preservation and revitalization of endangered minority language with use of speech technology tools

Maria Pendevska, Komercijalna Banka AD Skopje

Branislav Gerazov, Faculty of Electrical Engineering and Information Technology

Abstract Vanishing languages as under-resourced or low resourced languages are underrepresented on any media channel. One proximity of any knowledge is that only when it is used, it has and creates value. Our task is to find a way for this vanishing knowledge to be used. The Aromanian language faces both of these challenges. Even though it is officially recognized as a minority language in North Macedonia, the Aromanian language is under threat of disappearance. There are currently only a few thousand, mostly elderly speakers. One way to revive the language is to develop speech technology tools that will allow it to be included in digital media and applications, allowing easy access to the language by various age groups, especially the younger generations. In order to capture the knowledge of the language from the last generation of native speakers, we built a framework for digital presence of knowledge using Knowledge management practices. This framework integrates various new and old content, incorporating them in a logical structure. It contributes towards this goal by making the first necessary step - building a digital speech corpus for the endangered Aromanian language in North Macedonia. As a start we will include dialects spoken in our country, and in the future, we can scale the data with other dialects from neighboring countries, i.e. in the wider Balkan region. The data can then be used to develop text-to-speech synthesis and automatic speech recognition for Aromanian. These can be then used to facilitate language learning, especially among the younger "digital" generations. Moreover, the synthesizer can be used to enable the use of Augmented and Alternative Communication (AAC) devices, such as Communication Boards, for people with speech disabilities in Aromanian to be used worldwide, regardless of where they live.

Keywords: Under -resourced languages, Text-to speech technology and vice versa, Language digitalization

Evans, M. M., & Ali, N. (2013, October). Bridging knowledge management life cycle theory and practice. In *International Conference on Intellectual Capital, Knowledge Management and Organisational Learning ICICKM 2013–Conference Proceedings* (pp. 156-165).

Evans, M., Dalkir, K., & Bidian, C. (2014). A Holistic View of the Knowledge Life Cycle: The Knowledge Management Cycle (KMC) Model. *Electronic Journal of Knowledge Management 12*(2), 85–97.

Mittelmann A., Vollmar G., & and John U., Wissenmanagement-Kompetenzkatalog, Version 2.0, dated 20.02.2022 https://www.gfwm.de/wp-content/uploads/2017/04/GfWM_Kompetenzkatalog_WM_V2.0.pdf

Markovic M., HadjiLega H. J. & Trpeski D. (2019). Research on Aromanian language dialects and culture in North Macedonia, Macedonian Academy of Sciences and Arts (MANU) on Aromanian language and culture. http://aromanski.manu.edu.mk/

Pendevska, M. (2019). *The Influence of knowledge management on the innovation in the enterprises in Republic Macedonia*, (Doctoral dissertation, Defended June 2019 at Faculty of Economy, University Cyril and Methodius Skopje).

Translation

26

Very Lost in Translation: Overcoming OOV Challenges in Machine Translation for Low-Resource Languages through Loanwords Information

Felermino Ali, Univerty of Porto

Henrique Lopes Cardoso, Faculdade de Engenharia, Universidade do Porto / LIACC; Rui Sousa-Silva, Faculdade Letras, Universidade do Porto / CLUP

Neural Networks with Transformer-based architectures are currently the state-of-the-art for machine translation. However, they require a large amount of data which is often not available for many languages. One issue that contributes to the low performance of Neural Machine Translation (NMT) models for low-resource languages is the high prevalence of Out-Of-Vocabulary (OOV) words. This problem is particularly evident in predominantly spoken languages and is partly caused by the absence of standard spelling, particularly for loanwords – words borrowed from another language that enter the borrowing language with slight alterations. In this study, we propose methods to handle OOV resulting from loanwords. We consider signaling loanwords as input to the NMT model. These signaled tokens are subsequently used during postprocessing to translate the loanwords using a bilingual dictionary. Our research is grounded in the hypothesis that incorporating loanword information can significantly enhance the quality of translation for low-resource languages that have a significant amount of linguistic borrowing. To evaluate the effectiveness of our approach, we use a parallel corpus of Emakhuwa and Portuguese. Emakhuwa's has been chose for its low-resource nature and its significant lexical borrowing from Portuguese.

Keywords: Low-resource languages, Machine translation, Out-of-vocabulary

27

What we (all) do in the shadows: Language learning in the era of machine translation

Alice Delorme Benites, ZHAW Zurich University of Applied Sciences

Elizabeth Steele, Bfh Bern University of Applied Sciences; Sara Cotelli, University of Neuchâtel; Caroline Lehr, ZHAW Zurich University of Applied Sciences

The advent of neural machine translation (MT) has not only disrupted the translation professions but also the teaching and learning of foreign languages. This latter issue has long been unacknowledged, despite the efforts of individual researchers (e.g. Briggs 2018, Jiménez-Crespo 2017, Tsai 2020, Vinall & Hellmich 2021, Klimova et al. 2022) to deal with the elephant in the room. At a broader level, the concept of MT literacy (Bowker & Buitrago Ciro 2019) has emerged, but there has been a lack of knowledge about how MT is actually used, especially in relation to language learning. The popular media often claim that MT obviates the need to learn foreign languages, striking fear into the hearts of many language teachers. For these reasons, four Swiss universities have joined forces to investigate the practices, attitudes, and representations of learners in higher education regarding the use of MT. We will present the results of a multilingual wide-scale survey we conducted in 2021 among all Swiss universities (6,400 respondents). We found that, even though most respondents claimed to be using MT, almost all stated that they considered language learning to still be necessary and that MT was a useful tool for this purpose. We observed a multitude of practices, situations and contexts in which MT is being used in varied and creative ways, contradicting the simplistic claim that language learners would just copy-paste texts into the machine and pass off the output as their own work. However, we also discovered a widespread lack of reflection and critical thinking when using these tools, especially regarding data privacy and possible consequences in the case of a mistranslation. These findings have helped us to fine-tune an MT literacy concept for language learning and teaching purposes. Since 2022, we have carried out 20 interventions to language teachers in various Swiss universities to foster their own MT literacy and empower them in harnessing the potential of MT in their classrooms. This is an iterative process in that each intervention provides us with new insights into the evolving needs and the concerns of language teachers, allowing us to refine our underlying concept. After attending our interventions, the teachers are able to take on multiplier roles for their learners regarding MT literacy. While the launch of easily accessible text generation tools such as ChatGPT has further disrupted language learning and teaching, we are applying the lessons learnt from our project so far to go beyond MT literacy and work towards a concept of a more general artificial intelligence (AI) literacy as a response to the newly arising challenges.

Keywords: Machine translation literacy, AI literacy, Language teaching, Machine translation

Bowker, L., & Ciro, J. B. (2019). *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Publishing Limited.

Briggs, N. (2018). Neural machine translation tools in the language classroom: Students' use, perceptions, and analyses. *Jalt call journal, 14*(1), 2-24.

Jiménez-Crespo, M. A. (2017) The role of translation technologies in Spanish language learning. *Journal of Spanish Language Teaching*, *4*(2), 181-193.

Klimova, B., Pikhart, M., Benites, A. D., Lehr, C., & Sanchez-Stockhammer, C. (2023). Neural machine translation in foreign language teaching and learning: a systematic review. *Education and Information Technologies, 28*(1), 663-682.

Tsai, S. C. (2020) Chinese students' perceptions of using Google Translate as a translingual CALL tool in EFL writing. *Computer Assisted Language Learning, 35*(5-6), 1250-1272.

Vinall, K. & Hellmich, E.A. (2021). Down the rabbit hole: Machine translation, metaphor, and instructor identity and agency. *Second Language Research & Practice, 2*(1), 98-118.

28

Challenges for a Gender Inclusive Machine Translation

Beatrice Savoldi, Fondazione Bruno Kessler

Gender inclusivity in language has become a topic of central importance in contemporary discourse, with efforts being made to overcome gender bias and discrimination in various linguistic contexts (Motschenbacher, 2014; Comandini, 2021). Given the increasing reliance on tools that automatize language-based activities, such communicative contexts are no longer restricted to human practices only, and regard the communication supported by language technologies, too (Lauscher et al., 2021). Among them, machine translation (MT) -the task of automatically translating text across languages - was found to favor the generation of masculine forms and replicate stereotypical associations, e.g., doctors rendered as men, but nurses as women. While much research has been devoted to the topic of gender bias in MT (Savoldi et al., 2021), existing studies have focused on binary masculine-feminine forms to investigate - perhaps retroactively - the decisions incorporated in the design of language technology that lead to the emergence of bias. As we start to think about proactive measures toward the development of more inclusive technology, however, the dichotomous masculine/feminine approach appears limiting. Indeed, it neglects non-binary identities. Additionally, the feminine/masculine repertoire does not offer straightforward solutions to confront expressions that do not entail a single, correct translation. Such is the case of the English sentence "Doctors and nurses are needed", to be rendered in a grammatical gender language like Italian, which conveys gendered distinctions on several parts of speech (Corbett, 2013) e.g., It: Dottori/esse e infermieri/e sono richiesti/e. Towards alternative inclusive translation strategies for MT, it is from the realm of non-binary linguistic strategies that a path ahead can be found. Such a path, however, poses several (cross)linguistic, theoretical, and technical challenges: i) the evolution of non-binary linguistic strategies is ongoing and inconsistent across languages; ii) their incorporation in MT models might be perceived as a form of linguistic policing; iii) data, models and evaluation practices for the development of inclusive MT are lacking. In this presentation, I will lay out and systematize such challenges and present ongoing work to address them. Such work is of interest for LITHME, in particular in relation to the themes of linguistic ideologies and computational linguistics, since it emphasizes the impact of language technologies on our communicative practices and user's attitude towards the incorporation of such new, emerging practices in technology at scale.

Keywords: Machine Translation, Inclusive Language, Gender Bias

Comandini, G. (2021). Salve a tuttə, tutt*, tuttu, tuttx e tutt@: l'uso delle strategie di neutralizzazione di genere nella comunità queer online.: Indagine su un corpus di italiano scritto informale sul web. *Testo e Senso*, (23), 43-64.

Corbett, G.G. (2013). *The Expression of Gender*. De Gruyter.

Lauscher, A., Crowley, A., & Hovy, D. (2022). Welcome to the Modern World of Pronouns: Identity-Inclusive Natural Language Processing beyond Gender. In *Proceedings of COLING (pp. 1221-1232).*

Motschenbacher, H. (2014). Grammatical gender as a challenge for language policy: The (im)possibility of non-heteronormative language use in German versus English. *Language policy, 13*, 243-261.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics, 9*, 845-874.

29

Sustainable machine translation in global development: Establishing a framework for evaluating the sustainability of commercial multilingual NMT models

Matt Riemland, Dublin City University

Neural machine translation (NMT) has revolutionized interlingual communication for many of the world's language communities. In particular, multilingual NMT systems built and maintained by multinational technology corporations such as Google, Meta (Facebook), and Microsoft have dramatically expanded the range of (low-resource) languages available for automatic translation. These companies emphasize the technology's humanitarian impacts as well as their NMT systems' ethically responsible, sustainable designs (Doss Mohan and Skotdal, 2021; Bapna et al., 2022; Costa-jussà et al., 2022). However, sustainability claims made by private-sector humanitarian initiatives have drawn considerable scrutiny. Leaders of marginalized communities and Global South scholars have long contended that the dominant, neoliberal approach to global development constructs and applies a self-affirming notion of sustainability, concealing the prioritization of commercial profits over environmental and social concerns (Banerjee, 2003; Leal, 2007; Kari-Oca II Declaration of 2012). Furthermore, scholarly discussions of MT sustainability have largely focused on NMT in commercial contexts (Kenny et al., 2020). The alleged incompatibility between the private sector's profit-seeking imperative and genuine sustainability merits a critical, comprehensive framework for assessing MT sustainability in development settings. This presentation argues that, in order to gauge the technology's overall impact on marginalized peoples, researchers must evaluate MT sustainability in terms of three key dimensions: quality, social, and environmental. It posits these dimensions' applicability to the sustainability claims of the aforementioned commercial MT models. The quality dimension is arguably the preeminent factor in MT sustainability. Although commercial NMT providers are continually expanding the range of available languages, a large number of the added low-resource languages yield low-quality output, while researchers also urge "appropriate skepticism" toward their quality measurements and quality assessment methods (Bapna et al., 2022, p. 23). The social dimension is another important factor, as research demonstrates that poorly implemented information and communication technologies such as MT may even exacerbate social inequalities in development settings (Chipidza and Leidner, 2019). This second dimension entails, among other criteria, the implementation of comprehensive security measures for data, as well as the capacity for local communities to own and operate the technology autonomously. Free-to-use commercial NMT services are known to raise concerns in these areas. The environmental dimension is the final component of MT sustainability. In response to widespread criticisms of the technology's massive carbon footprints, Google, Meta, and Microsoft claim to offset the emissions of their massive multilingual NMT systems through renewable energy purchasing (Patterson et al., 2022, p. 24). Nevertheless, scholars have asserted flaws in such self-reported sustainability metrics and carbon accounting methods from multinational corporations (see Brander et al., 2018; Chiapello and Engels, 2021; Klaaßen and Stoll, 2021). Given the stark Digital Divide between high- and low-resource languages, the default ideological position may construe the addition of low-resource languages to massive multilingual NMT models as an unambiguously positive step toward language vitality and digital inclusion. This presentation seeks to illuminate the ethical complexities and wide-ranging impacts of such developments. It suggests empirical research opportunities for interrogating multinational technology companies' claims to sustainability for each of the proposed dimensions.

Keywords: Low-resource languages, MT sustainability, Multilingual NMT, MT ethics

Alejandro Leal, P. A. (2007). Participation: The ascendancy of a buzzword in the neo-liberal era. *Development in Practice*, *17*(4–5), 539–548. https://doi.org/10.1080/09614520701469518

Banerjee, S. B. (2003). Who Sustains Whose Development? Sustainable Development and the Reinvention of Nature. *Organization Studies, 24*(1), 143–180.

Bapna, A., Caswell, I., Kreutzer, J., Firat, O., van Esch, D., Siddhant, A., Niu, M., Baljekar, P., Garcia, X., Macherey, W., Breiner, T., Axelrod, V., Riesa, J., Cao, Y., Chen, M. X., Macherey, K., Krikun, M., Wang, P., Gutkin, A., … Hughes, M. (2022). Building Machine Translation Systems for the Next Thousand Languages *arXiv preprint arXiv:2205.03983*. Google. https://doi.org/10.48550/arXiv.2205.03983

Brander, M., Gillenwater, M., & Ascui, F. (2018). Creative accounting: A critical perspective on the market-based method for reporting purchased electricity (scope 2) emissions. *Energy Policy, 112*, 29–33.

Chiapello, E., & Engels, A. (2021). The fabrication of environmental intangibles as a questionable response to environmental problems. *Journal of Cultural Economy, 14*(5), 517–532.

Chipidza, W., & Leidner, D. (2019). A review of the ICT-enabled development literature: Towards a power parity theory of ICT4D. *The Journal of Strategic Information Systems, 28*(2), 145–174.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., … Wang, J. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation *arXiv preprint arXiv:2207.04672*.

Doss Mohan, K., & Skotdal, J. (2021, October 11). Microsoft Translator: Now translating 100 languages and counting! Microsoft Research. https://www.microsoft.com/en-us/research/blog/microsoft-translator-now-translating-100-languages-and-counting/

Kari-Oca II Declaration. (2012, June 17). Indigenous Peoples Global Conference on Rio+20 and Mother Earth. Indigenous peoples global conference on Rio+20 and mother earth, Kari-Oka Village, at Sacred Kari-Oka Púku, Rio de Janeiro, Brazil. https://www.ienearth.org/kari-oca-2-declaration/

Klaaßen, L., & Stoll, C. (2021). Harmonizing corporate carbon footprints. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-26349-x

Kenny, D., Moorkens, J., & do Carmo, F. (2020). Fair MT: Towards ethical, sustainable machine translation. *Translation Spaces*, *9*(1), 1–11. https://doi.org/10.1075/ts.00018.int

Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L. M., ... & Dean, J. (2022). The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. *Computer*, *55*(7), 18–28. https://doi.org/10.1109/MC.2022.3148714

30

What is Machine Translation User Experience and why should we start looking at it? An overview in multilingual communication processes

Vicent Briva-Iglesias, Dublin City University

Recent advances in language technologies have grown exponentially in recent years, and the impact that new technologies are having on our society, the way we relate, work, or interact with new systems or technological products is changing faster and faster. Machine Translation (MT) is one of the elements that has had the greatest impact recently, especially in multilingual communication processes. Consequently, the spectrum of MT users is growing: professional translators, people with no knowledge of a language who need to understand the content of a text, academics writing in a language that is not their mother tongue, or even people in crisis situations such as earthquakes or wars. Despite the vital importance of MT, the main focus has been on the productivity and quality offered by the adoption of such systems. In the intrinsic human-machine interactions that occur when using MT, there is one element that has been often overlooked and neglected in research and industry: the user. What does any MT user feel or experience when interacting with MT systems or products? Is this experience appropriate to what the user is looking for? Do developers of new products take users' needs into account? Or, alternatively, are changes and improvements in MT systems only technical and not socio-technical? Through a transdisciplinary study, drawing elements from Translation Studies, Multilingual Communication, Human-Computer Interaction and Machine Translation, we propose the Machine Translation User Experience (MTUX) concept as a tool to measure the user experience when interacting with MT. We present the results of a project funded by the European Association for Machine Translation in which we analyse how 15 professional translators feel when interacting with MT, and develop a methodology to measure these user experiences, which can be applied to any kind of MT user. The ultimate goal is to consider MTUX as a key element in new language technology development. MTUX is a forgotten element in MT research to date. Nevertheless, its analysis is of great relevance for discovering users' pain points in their interaction with MT. Knowing these weak points of the systems will allow us to introduce changes and personalise the experience for each type of user and thus reduce the rejection of new language technologies by means of a better MTUX. People working and interacting with MT will enjoy better experiences if MTUX is considered, which may translate huge benefits in many different domains. Some examples are happier language professionals in the language services industry, who will offer higher quality texts, and ultimately happier customers and language service providers. Also, people in crisis scenarios will be able to receive vital assistance in minor languages at higher speed through appropriate means and according to their needs. MTUX is therefore something we should start looking at in multilingual communication processes.

Keywords: Machine translation, User experience, Human-computer interaction

31 Paidiom: a text preprocessing algorithm to improve the neural machine translation of multiword expressions

Carlos Manuel Hidalgo Ternero, University of Malaga

The recent emergence of neural networks in machine translation has represented a real breakthrough, bringing forth Neural Machine Translation (NMT), which has resulted in a considerable qualitative leap compared to previous ruled-based and statistical models (Bentivogli et al., 2016; Junczys-Dowmunt et al., 2016; Wang et al., 2022). Despite these advances, NMT systems still have an important weak point: multiword expressions (MWEs). Besides their quintessential problematic features such as syntactic anomaly, non-compositionality, diasystematic variation and ambiguity, among others, a further challenge arises for NMT: MWEs do not always consist of adjacent tokens (e.g., "They took my proposal into consideration."), which seriously hinders their automatic detection and translation (Constant et al., 2017; Corpas Pastor, 2013; Monti et al., 2018; Ramisch & Villavicencio, 2018; Rohanian et al., 2019). To overcome the challenges that discontinuous MWEs still pose for NMT (cf. Colson, 2019; Zaninello & Birch, 2020), we have designed an upgraded algorithm, called Paidiom, that is able not only to automatically convert discontinuous MWEs into their continuous form (analogously to our previous algorithm gApp) but also to translemmatise them, i.e., to directly convert MWEs into their target-text equivalents in order to improve NMT; thus highly contributing to one of LITHME's central themes. To test Paidiom's effectiveness, the performance of VIP (cf. Corpas Pastor, 2021), Google Translate and DeepL's NMT systems was examined against a total of 400 cases, comprising 100 discontinuous forms (i.e., the original texts), 100 continuous forms after gApp's conversion, 100 continuous and translemmatised forms after Paidiom's conversion, and 100 continuous and translemmatised forms after the manual conversion (our gold standard) of the Spanish MWEs "haber gato encerrado" ('there is something fishy going on'), "ser cuatro gatos" ('to be a small bunch of people'), "dormir la mona" ('to sleep something off'), and "ganar/costar/pagar cuatro perras" ('to earn/cost/pay peanuts'), in the ES>EN translation direction. For this text set, the present experiment yielded promising global results that, on the one hand, go in line with our previous experiments with gApp (Hidalgo-Ternero, 2021 and 2023; Hidalgo-Ternero & Corpas Pastor, 2020, 2023a and 2023b; Hidalgo-Ternero, Lista & Corpas Pastor, 2023, and Hidalgo-Ternero & Zhou-Lian, 2022 and 2023), along which we proved that NMT of discontinuous MWEs can overall be improved by converting them into their continuous form: in this experiment, NMT systems achieved an 8.3% accuracy in the discontinuous form vs. 13% in the continuous form after gApp, i.e., an enhancement by 4.7%. On the other hand, global results also confirmed our initial hypothesis: NMT systems can deliver a considerably better performance if MWEs are not only converted into their continuous form but also translemmatised prior to NMT. In the light of these results, MWEs' conversion into their continuous form plus their translemmatisation with Paidiom led NMT systems to achieve an overall 91.7% accuracy, attaining an analogous performance to the gold standard (92.7%). When contrasted with the original (discontinuous) version, Paidiom could, on average, improve NMT by 83.4% (84.4% with the gold standard).

Keywords: Neural machine translation, Multiword expressions, Text preprocessing algorithm, Spanish, English

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2018). Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.

Colson, J. P. (2019). Multi-word Units in machine translation: why the tip of the iceberg remains problematic – and a tentative corpus-driven solution. [Conference presentation] MUMTT2019.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, *43*(4), 1–92.

Corpas Pastor, G. (2013). Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. In I. Olza & E. Manero (Eds.) *Fraseopragmática. Colección Romanistik* (pp. 335-373). Frank & Timme.

Corpas Pastor, G. (2021). Technology Solutions for Interpreters: The VIP System. *Hermēneus. Revista de Traducción e Interpretación*, *23*, 91-123.

Hidalgo-Ternero, C. M. (2021). El algoritmo ReGap para la mejora de la traducción automática neuronal de expresiones pluriverbales discontinuas (FR>EN/ES). In G. Corpas Pastor, M. R. Bautista Zambrana & C. M. Hidalgo-Ternero (Eds.), *Sistemas fraseológicos en contraste: enfoques computacionales y de corpus (pp. 253-270)*. Editorial Comares.

Hidalgo-Ternero C. M. (2023/forthcoming). A la cabeza de la traducción automática neuronal asistida por gApp: somatismos en VIP, DeepL y Google Translate. In *G. Corpas Pastor y M. Seghiri (Eds.), Aplicaciones didácticas de las tecnologías de la interpretación*. Comares.

Hidalgo-Ternero, C. M., & Corpas Pastor, G. (2020). Bridging the "gApp": improving neural machine translation systems for multiword expression detection. *Yearbook of Phraseology*, *11*, 61-80. https://doi.org/10.1515/phras-2020-0005

Hidalgo-Ternero C. M., & Corpas Pastor, G. (2023a/forthcoming). Qué se traerá gApp entre manos… O cómo mejorar la traducción automática neuronal de variantes somáticas (ES>EN/DE/FR/IT/PT). In Seghiri, M. & Pérez Carrasco, M. (Eds.). *Aproximación a la traducción especializada*. Peter Lang.

Hidalgo-Ternero C. M., & Corpas Pastor, G. (2023b/forthcoming). ReGap: a text preprocessing algorithm to enhance MWE-aware neural machine translation systems. In J. Monti, G. Corpas Pastor y R. Mitkov (Eds.), *Recent Advances in MWU in Machine Translation and Translation technology*. John Benjamins Publishing Company.

Hidalgo-Ternero C. M., Lista, F. & Corpas Pastor, G. (2023/under review). gApp-assisted NMT: how to improve the neural machine translation of discontinuous multiword expressions (IT>EN/DE). Language Resources and Evaluation.

Hidalgo-Ternero, C. M., & Zhou-Lian, X. (2022). Reassessing gApp: does MWE discontinuity always pose a challenge to Neural Machine Translation? In G. Corpas Pastor y R. Mitkov (eds.), *Computational and Corpus-Based Phraseology* (pp. 116–132). Springer.

Hidalgo-Ternero, C. M., & Zhou-Lian, X. (2023/under review). Minding the gApp in the ES>EN/ZH neural machine translation of discontinuous multiword expressions. Natural Language Engineering.

Junczys-Dowmunt, M., Dwojak, T. & Hoang, H. (2016). Is neural machine translation ready for deployment? A case study on 30 translation directions. *arXiv preprint arXiv:1610.01108*.

Monti, J., Seretan, V., Corpas Pastor, G., & Mitkov, R. (2018). Multiword units in machine translation and technology. In *R.* Mitkov, J. Monti, G. Corpas Pastor & V. Seretan (Eds.), *Multiword Units in Translation and Translation Technology* (pp. 1-37). John Benjamins.

Ramisch, C., & Villavicencio, A. (2018). Computational treatment of multiword expressions. In R. Mitkov (Ed.), *Oxford Handbook on Computational Linguistics* (2ª ed).

Rohanian, O., Taslimipoor, S., Kouchaki, S., An Ha, L., & Mitkov, R. (2019). Bridging the Gap: Attending to Discontinuity in Identification of Multiword Expressions. In J. Burstein, C. Doran, & T. Solorio (Eds.), 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1, 2692–2698.

Wang, H., Wu, H. He, Z., Huang, L. & Church, K. W. (2022) Progress in Machine Translation, *18*, 143-153. *Engineering.*

Zaninello, A., & Birch, A. (2020, May). Multiword expression aware neural machine translation. In Proceedings of the 12th Conference on Language Resources and Evaluation (pp. 3816–3825).

Tools

33 Document Evaluation with LegalTech Tools

Miklós Zorkóczy

LITHME relevance: How LegalTech can support the document evaluation work in the human – machine era? LegalTech: A document evaluation system is a LegalTech tool. But what is LegalTech? A LegalTech tool is developed to support the legal operation. Such support can be document automation, legal chatbot communication, legal research, online dispute resolution systems, contract lifecycle management systems, case management systems, and last but not least document evaluation systems. Supported legal tasks: What legal tasks can be supported by a document evaluating machine? Like preparing for an audit, creating valuation in an M&A transaction, evaluating court files, investigate invoices in tax issues. Can a machine understand the legal text? How can you train a machine without coding skills? Mostly the systems available on the market don't understand the language, the text, the content, the context. They don't know the semantic relation. They only recognize characters. Characters are data. Data can be classified, structured, and calculated with data science tools. A system can be trained to do it in a very cheap, fast, and effective way. Machine learning: Machines are very fast and effective to find patterns set by humans. Given the patterns to the system, it classifies documents into categories. So, it can structure the database, put the files in directories. All the corporate, litigation, labour, tax, data privacy documents are vetted and classified. Why is it important? Imagine a group of lawyers, who work in practice groups. Like corporate practice group, labour law practice group, French speaking practice group, EU competition law practice group. With the help of the machine, you can delegate tasks easily to practice groups. It is not necessary to spend human chargeable hours on reading all the documents and classify them one by one. It does not matter for the machine, if there are 100s or 1000s of documents, the speed and efficiency of the machine work will be on an incredibly higher level, than humans'. This is time and cost saving. Document evaluation: Next step for the machine is to filter the content. In a structured database, you can filter the findings given to the machine. Like, if I wanted to deliver a data privacy audit, I need to find all the data protection related clauses. I need to search for clauses mentioning data subject's rights, personal data, sensitive data, data processor, data controller. These are the patterns. And the machine will list all the findings. So the next step is to open the files on the list, what the machine had searched for you upon your instructions. Finally, you need to check the findings of the machine if the system was right with the findings or not. As a result of the work, you will have an extract in an excel, or word file. This would be an extract of the due diligence report. Why is it important to attach the extract? The extract proves and demonstrates all of your comments, statements and findings in your final due diligence report. This system is not replacing but enhancing lawyers' performance. The presentation will discuss the legal and technology aspects of such systems.

Keywords: LegalTech, Machine Learning, Explainable AI, Document Evaluation, Legal Operation

Bassli, L. (2020). *Simple guide to legal innovation*. American Bar Association.

Békés, G., Kézdi, G. (2021). *Data analysis for business, economics, and policy*. Cambridge University Press.

Bhatti, S.A., Datoo, A., Chishti, S., Indjic, D. Dr., (2020). *The LegalTech Book*. Wiley.

Susskind, R. (2017). *Tomorrow's lawyers: An introduction to your future*. Oxford University Press.

Ződi, Zs. (2012). Legal database and legal research. Gondolat.

34 Online translation resources in the process of human translation (HT) and the post-editing of machine translation (MTPE)

Eszter Sermann, University of Szeged, Hungary

Research has confirmed that the use of online translation resources constitutes a considerable proportion of the overall time spent on translation (Hvelplund, 2017); moreover, with the constant technological changes in recent years, increasing attention within translation process-oriented research has been given to the types of translation resources and their use in the process of human translation and post-editing of machine translation (e.g. Gough, 2017; Hvelplund, 2011, 2016, 2017; Prieto Ramos, 2021). The quality of translations/post-edited texts is often expressed in error numbers: Čulo and Nitzke (2016) found that HT texts were characterized by more accurate terminology use than MTPE texts. In the spring of 2020, the Translation and MT Post-Editing Competence Research Group of the University of Szeged, Hungary started a comprehensive study in order to reveal what role PACTE's translation sub-competences play in human translation and post-editing of machine translation. In this paper, we will present the results of a process-oriented research carried out in order to study the types of translation resources used by translator trainees during human translation and post-editing of the same legal text. Data collection was carried out in the spring and fall of 2022, during which 14 first-year and 12 second-year master's students of translation translated and post-edited a part of a copyright agreement from English into Hungarian. (The MT output for the post-editors was produced by eTranslation, the machine translation tool of the EU). Based on the screen videos of students' workflow (recorded with the OBS Studio software), the types of online translation resources (monolingual, bilingual dictionaries, corpus-based dictionaries, terminology databases, machine translation engines, Google search, legal documents) were identified, and the data on translators and post-editors were compared. Afterwards, 10 key terms were selected from the source text, and their equivalents were evaluated in translated and post-edited target language texts. The expected results will show (1) the types of resources used by translators and post-editors; (2) the similarities and differences between the two groups studied; and (3) whether there is a relation between the types of resources used, the mode of translation (HT or MTPE) and the adequacy of the target language equivalents of legal terms. The research presented in this paper is closely related to the following themes and aims of LITHME WG2: machine translation of legal texts, liability for machine translation, problems of equivalence in legal translation, language technologies in translation, blurring borders between human and machine translation.

Keywords: Process-oriented research, Legal translation, Online translation resources, Human translation, Post-editing of machine translation

Čulo, O. & Nitzke, J. (2016). Patterns of terminological variation in post-editing and of cognate use in machine translation in contrast to human translation. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation* (pp. 106–114) https://aclanthology.org/W16-3401

Gough, J. (2017). Investigating the use of resources in the translation process. In Pastor, G. C. & Durán Muñoz, I. (eds.) *Trends in e-tools and resources for translators and interpreters* (Vol. 45) Brill (pp. 9–36).

Hvelplund, K. T. (2011). *Allocation of Cognitive Resources in Translation: An eye-tracking and Key-log-ging Study*. [PhD thesis] Copenhagen Business School, https://www.econstor.eu/bitstream/10419/208778/1/cbs-phd2011-10.pdf

Hvelplund, K. T. (2016). Cognitive efficiency in translation. In Muñoz Martín, R. (ed.) *Reembedding Translation Process Research, Amsterdam and Philadelphia: John Benjamins* (pp. 149–170)

Hvelplund, K. T. (2017). Translators' use of digital resources during translation. *HERMES–Journal of Language and Communication in Business*, (56), 71–87, https://doi.org/10.7146/hjlcb.v0i56.97205

Prieto Ramos, P. (2021) The use of resources for legal terminological decision-making: patterns and profile variations among institutional translators. P*erspectives*, *29(*2), 278–310, DOI: 10.1080/0907676X.2020.1803376

35

New technological tools in audiovisual translation – pitfalls and benefits

Judit Sereg, ELTE Eötvös Loránd Science University

Audiovisual translation has always been a field of translation closely connected with language technology. However, in the past, the technological aspect of audiovisual translation had been considered mainly in terms of constraints (snychronization in dubbing, spatial constraints in subtitling), and very little thought has been given to the benefits of the use of language technology. In the last decade, and especially in the last few years, new technological tools are appearing in audiovisual translation. Deepfake dubbing makes it possible to change the movements of an actor's lips to match that of the dubbing text, alleviating the need to write words that might not be natural in the target language, but match the original actor's lip movements. It can even change the pitch of the dubbing actor's voice to match it more closely to that of the original. In the meantime, deepfake technology can also be seen as a method for manipulation as it "can generate […] a humorous, pornographic, or political video of a person saying anything, without the consent of the person whose image and voice is involved" (Westerlund 2019). Machine translation, especially in the case of subtitles is getting more widely used, since it speeds up the translation process and a larger quantity of audiovisual content can be localized in a smaller amount of time. Still, if the post-editing process is not executed properly, the low quality of subtitles can have a negative effect on the ratings of programs, and diminish the entertainment or informational value for the target audience. The use of machine translation for dubbing purposes raises even more issues, as dubbing's "aim is to create the illusion that the onscreen characters are speaking in the target language, i.e. in the language of the audience" (Baumgarten 2005: 21). In countries with less widely spoken languages, dubbed products can be the main source of authentic language use, and if the quality of this language is poor, it might lead to the overall decline in the diversity of that language. Automatically generated subtitles and voice-overs might make it simple for companies to adhere to accessibility rules, but without proper quality control, disadvantaged groups might have to accept content with inferior quality and less usability. In my presentation, I would like to exhibit the various new technological tools directly connected to audiovisual translation and showcase what kind of issues need to be addressed and taken into account when working with them, be it in the daily work of the translator or at the level of a dubbing studio or streaming service.

Keywords: Audiovisual translation, New technologies, Dubbing, Deep fake, Subtitling

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11), 40-53. http://doi.org/10.22215/timreview/1282

Baumgarten, N. (2005). *The Secret Agent: Film dubbing and the influence of the English language on German communicative preferences. Towards a model for the analysis of language use in visual media.* [PhD Dissertation.] Universität Hamburg. DOI: 10.13140/RG.2.2.19258.93126

Power

36

"I'm straight!" Challenging content moderation on (Slovak) Facebook

Lucia Molnár Satinská, Slovak Academy of Sciences

Roman Soóky, Slovak Academy of Sciences

Two people were killed and one severely injured in front of a queer bar in October 2022 in Bratislava, the capital of Slovakia, which was proved to be a hate crime against the LGBTQ+ community. The deed motivated a wave of solidarity, but also initiated hate speech statements both from the general public and some prominent figures. The presentation aims to uncover the communication strategies in their statements which include hate speech undetectable by social media protection algorithms. It thus aims to display how ineffective content moderation on Facebook is in the Slovak language. For example, some (populist) politicians started to employ phrases such as 'I'm straight!' in their communication on Facebook. Although, one cannot ultimately deem such expressions offensive or hate speech. However, an analysis of the politicians' message may suggest the opposite (e.g. Soóky 2022), taking into notice the socio-political context and framing (especially its timing and fake solidarity – i.e. solidarity with a majority whose rights are not violated). In the context of commercial content moderation of platforms like Facebook (i.e. Meta), which is supposed to remove any material considered harmful or offensive from user-generated content (Lau 2022), the actual Facebook content–moderation scheme is failing to reach its promises (Patel & Hecht-Felella 2021). When this concerns a less-resourced language (such as Slovak), the more pressing the issue gets. Moderation on Facebook fails to restrict utterly harmful content against the LGBTQ+ community in politicians' public communication neglecting their influential potentiality as public figures. The troubling content remains visible and spreadable, albeit the Community Standards on Facebook view hate speech as restricted content (Facebook 2023), leading to the radicalization of its users, which might result in further violence and hate crimes. The presentation deals with dilemmas concerning hate speech regulations on Facebook with examples from the Slovak context and presents some possible ways of how to overcome such issues.

Keywords: Language policy, Content moderation, Hate speech, Communication strategies, LGBTQ+

Facebook Community Standards. Available at: https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/#policy-details [Last access 23-02-2023].

Lau, M. (2022). Content moderation as language policy: Connecting commercial content moderation policies, regulations, and language policy. *Working papers in Applied Linguistics and Linguistics at York*, *2*, 1-12.

Patel, F., & Hecht-Felella, L. (2021). Oversight Board's First Rulings Show Facebook's Rules Are a Mess. *Just Security*.

Soóky, R. (2022). How (un)freedom is being developed: The linguistic world of one political party in Slovakia concerning liberalism. *Jazykovedný časopis*, *73*(3), 329-353.

37

The privatisation of language: Commercial language management and its effects on standard language ideologies

Iker Erdocia, Dublin City University

Bettina Migge, University College Dublin; Britta Schneider, Europa-Universität Viadrina; Helen Kelly-Holmes, University of Limerick

This paper explores the effects of digital technologies on standard language ideologies, specifically in relation to gender-neutral language (GNL). For centuries, language academies, academics, governmental institutions or, in some cases, publishers, have been entrusted with the role of serving as national linguistic authorities. These bodies uphold and enforce standard languages while stigmatising diverse, everyday practices that do not conform to strict standards. However, in the last few decades there have been public debates on language norms, particularly in regard to GNL forms. Proposals put forth by social activists and political organisations have in some contexts encountered opposition from language academies and standardising institutions. Some of these institutions contend, for example, that the use of masculine generics as the unmarked form encompasses both masculine and feminine grammatical genders. The debate on GNL is a complex and heated one, often characterised by ideological and political connotations (Erdocia 2022). In an era of late modern capitalist tech culture (Schneider 2022), technological corporations are becoming relevant actors in the enterprise of language standardisation. They are reorganising not only the ways in which we use language (Kelly-Holmes 2022) but also how language itself is managed. Some AI language models (e.g. Google Docs, Chat GPT) seem to be programmed to a certain extent to use inclusive and non-sexist language. For example, the question to Chat GPT regarding its use of GNL options prompts the response from the model that the use of non-sexist language is becoming common and that it is 'an important part of the fight for gender equality and can help to create a more inclusive and just society'. Microsoft Word's 'Editor' function has an 'Inclusiveness' check, which will suggest changing terms like 'fireman' to 'firefighter', with the comment that 'a gender-neutral term would be more inclusive'. In reaction to this trend in the digital realm, there are attempts from language academies to linguistically discipline technological companies by reproducing the monopoly of language authority in cyberspace. A noteworthy instance of this phenomenon is the case of the Spanish language. The Spanish Academy recently announced its goal to establish linguistic regulations for AI and is currently seeking public funding to do so. It is unclear whether the Spanish Government, including some ministers who openly use GNL forms, will support such an initiative. This case most interestingly shows that public or public-funded institutions, such as language academies, may not be in a position to reproduce their role as national linguistic authorities. Private companies offering language technologies may challenge language regulatory practices in public space. Drawing on these cases, we discuss how the transition to a privatised, fragmented and deregulated production and consumption of language technology products is destabilising traditional hierarchical orders in language management, which may at the same time be regarded as indicating a trend towards the fragmentation of public space. We conclude by outlining the main tenets of the concept of privatisation of language and considering ways in which commercial companies, institutional actors and users could engage in a more open dialogue.

Keywords: Language ideologies, Language technologies, Language management, Language in cyberspace, Gender-neutral language

Erdocia, I. (2022). Language and culture wars: The far right's struggle against gender-neutral language. *Journal of Language and Politics*, *21*(6), 847-866.

Kelly-Holmes, H. (2022). Sociolinguistics in an increasingly technologized reality. *Sociolinguistica*, *36*(1-2), 99-110.

Schneider, B. (2022). Multilingualism and AI: The regimentation of language in the age of digital capitalism. *Signs and Society*, *10*(3), 362-387.

38 Blockly-ed Dialogue: Empowerment or Reinforcement of Power?

Sviatlana Höhn, LuxAI S.A.

Nina Hosseini-Kivanani, University of Luxembourg

Researchers working on interaction often rely on cooperation with technical specialists who provide the necessary technical basis and support for their research questions related to social interaction, pragmatics, and similar topics (e.g., Dippold (2023)). At the same time, block-based programming libraries, such as Blockly, have been shown to empower and motivate groups underrepresented in technological professions to become interested in coding (Seraj et al. 2019). Several attempts to use Blockly programming for industrial robots (Winterer et al. 2020) and education (Rahaman et al. 2020) have led to the conclusion that Blockly can be used to write large and complex robot programs that are easily readable, understandable, and maintainable. Some social robotics platforms provide access to block-based robot programming, enabling non-technical specialists such as autism therapists to create their own applications for human-robot interaction (Costa et al. 2017). In this talk, we will look at one such platform called QTrobot Studio (https://docs.luxai.com/docs/intro_graphical) from the perspective of language ideologies, language rights, language learning, and interaction. QTrobot Studio empowers and supports therapists in the code-free, Blockly-based creation of robot-assisted autism therapy (RAAT) applications for children. The interactions for this use-case need to be prototypical, and the minimally-expressive type of interaction helps children with autism to acquire language, social, emotional, and cognitive skills. However, even for such simplified interactions, a combination of various language and robotic technologies is needed, including text-to-speech (TTS) and speech recognition, face recognition, posture tracking, and so on. TTS for QTrobot Studio is provided by companies specializing in speech technology, making it easier to integrate new languages, but also limiting the choices to those provided by 3rd parties. Currently, QTrobot Studio offers TTS in 27 languages, and speech recognition can be selected from only seven languages, two of which are variations of English. All those languages of TTS and speech recognition are standard languages associated with nation states. Language variations associated with social closeness, family, belonging and intimacy are not offered. The creation of RAAT applications using TTS in the language of the child's parents is possible if parents speak one of those (standard) languages. For all other languages and variations, all utterances need to be pre-recorded and played as sound recordings. However, the latter makes authoring and maintenance more complex: for any tiny change in an utterance, a new voice recording has to be created. As a consequence, families speaking one of the prestigious languages receive better care than families who speak marginalized languages. In this way, technology reinforces social hierarchies of languages, even though technology could play an important role in destigmatizing minoritized languages and facilitating language rights, as stated in (de-Dios-Flores et al. 2022). Overall, QTrobot Studio offers an empowering way of code-free interaction design. Nevertheless, it is essential to be aware of cross-disciplinary issues such as the reinforcement of language ideologies, social language hierarchies and their consequences for language learning opportunities for vulnerable populations. We will make suggestions on how to solve them.

Keywords: Code-free programming, QTrobot Studio, Speech Technology, Language Standards, Language Rights

Costa, A. P., Steffgen, G., Lera, F. R., Nazarikhorram, A., & Ziafati, P. (2017, March). Socially assistive robots for teaching emotional abilities to children with autism spectrum disorder. In *3rd Workshop on Child-Robot Interaction at HRI*.

de-Dios-Flores, I., Magarinos, C., Vladu, A. I., Ortega, J. E., Campos, J. R. P., Garcia, M., ... & Regueira, X. L. (2022, June). The Nós Project: Opening routes for the Galician language in the field of language technologies. In *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference* (pp. 52-61).

Dippold, D. (2023). "Can I have the scan on Tuesday?" User repair in interaction with a task-oriented chatbot and the question of communication skills for AI. *Journal of Pragmatics*, *204*, 21-32.

Seraj, M., Katterfeldt, E. S., Bub, K., Autexier, S., & Drechsler, R. (2019, November). Scratch and Google Blockly: How girls' programming skills and attitudes are influenced. In *Proceedings of the 19th Koli Calling International Conference on Computing Education Research* (pp. 1-10).

Winterer, M., Salomon, C., Köberle, J., Ramler, R., & Schittengruber, M. (2020, September). An expert review on the applicability of Blockly for industrial robot programming. In *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)* (Vol. 1, pp. 1231-1234). IEEE.

Rahaman, M. M., Mahfuj, E., Haque, M. M., Shekdar, R. S., & Islam, K. Z. (2020). Educational robot for learning programming through Blockly based mobile application. *Journal of Technological Science & Engineering (JTSE)*, *1*(2), 21-25.

Interaction

39 Interactive Alignment in Human-Computer Interaction (HCI)

Dr. Netaya Lotze, University of Münster

We understand interactive alignment as a psycho-linguistic model for interpreting persistence in dialogue (Pickering & Garrod, 2004; Szmrecanyi, 2005; Koulouri et al., 2016). During language production, people tend to adapt to the previously received turn of their interlocutor; i.e. they adopt lexical or syntactic structures of the previous turn to produce their own turn. The phenomenon can be described as the human tendency to repeat a recently encountered structure. Alignment also has been found in Human-Computer Interaction (HCI) (Fischer, 2006; Lotze, 2016; Branigan & Pearson, 2016; Huiyang & Min, 2022), but its function is still controversial. We can either interpret it as a primal function of human memory and therefore as a pre-conscious priming effect (as in human-human communication), or as a conscious user strategy that aims to adapt to keywords currently used by the system in order to avoid disruptions in dialogue. In my talk, I am going to present two studies, in which we qualitatively analyse and quantify interactive alignment in HCI on a functional level: Lotze (2016) on chatbots and adult users and Lotze (in preparation) on chatbots and children at the age of 6-10 years. It is a mixed-method approach of both qualitative and quantitative analyses: conversational analysis (CA) and corpus linguistics. On a quantitative level, we can show that alignment in computer-mediated communication among humans is already much rarer than in oral communication, and even rarer in interactions with chatbots. Children attend more to interactive alignment than adults. I am going to show another recent study by my workgroup "AI + Language" (University of Münster, Germany), which deals with interactive alignment towards voice user interfaces (VUI), namely Amazon Alexa (Frommherz in preparation), as work in progress. We estimate that we can also find more interactive alignment in oral HCI with VUIs than in written contexts. Our work is colsely related to the aims of LITHME, because we apply psycho-linguistics on HCI in order to gain a deeper understanding oft the pragmatics of human-machine interaction as a currently rising form of dialogicity that will have a huge influence on society as a whole.

Keywords: Pragmatics of HCI, Interactive alignment, Psycho-linguistics, Corpus-linguistics, Conversational analysis

Huiyang, S., & Min, W. (2022). Improving interaction experience through lexical convergence: The prosocial effect of lexical alignment in human-human and human-computer interactions. *International Journal of Human-Computer Interaction*, *38*(1), 28–41.

Branigan, H., & Pearson, J. (2016). Alignment in human-computer interaction. In *K. Fischer (Ed.), Proceedings of the Workshop on How people talk to computers, robots, and other artificial communication partners* (pp. 140–156). Hanse-Wissenschaftskolleg Institute for Advanced Study.

Koulouri, T., Lauria, S., & Macredie, R. D. (2016). Do (and say) as I say: Linguistic adaptation in human–computer dialogs. *Human-Computer Interaction*, *31*(1), 59–95.

Lotze, N. (2016). *Chatbots: eine linguistische Analyse* (p. 443). Peter Lang International Academic Publishers.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190.

Szmrecsanyi, B. (2005). Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory*, *1*(1), 113–149.

40 Chatbots in public administration

Dimitra Anastasiou, Luxembourg Institute of Science and Technology

"You are being transferred to customer support": This is often the answer of a chatbot in customer service, when it cannot match the appropriate answer to the user query. What about chatbots used in public administration in Europe to facilitate administrative procedures, such as ID requests, taxation, asylum, etc.? In which languages are chat or voicebots in public administration available? In some European countries, such as Denmark, Estonia, Latvia, there are chatbots used in many public authorities, whereas in other countries, such as Romania or Luxembourg, there are not. In our talk, we will give some examples of chatbots used in public administration in Europe, which are linked with many challenges, such as large number of relevant services, the complexity of administrative services, the context-dependent relevance of user questions, the differences in expert-language and user-language as well as the necessity of providing highly reliable answers for all questions (Lommatzsch, 2018). To these reasons, we should add the language diversity in Europe. This language diversity has a consequence that each EU country and each administration uses its own initiative to deploy a chatbot (often monolingual) resulting in a scenario where the interaction with e-government through virtual assistants is scarce and fragmented. Multilingual bots and guides on how to create them are coming up increasingly in the last years (Boonstra, 2021), but also mainly by the industry and their business solutions. Many multilingual bots are used for foreign language learning, such as Mondly (supporting 41 languages) and Tutor Mike chatbot (learning English as a second language) to evaluate responses from a multilingual chatbot to determine the potential effectiveness. In the CEF funded project ENRICH4ALL, we integrated eTranslation into a commercial chatbot platform called BotStudio developed by the Danish SME SupWiz and deployed it into three public authorities in Luxembourg, Romania and Denmark. eTranslation is a CEF building block that can be integrated into digital services to add translation capabilities. It currently covers not only the 24 official languages of the EU, but also Ukranian, Russian, simplified Chinese, Turkish, and Arabic. This work relates to the LIHTME because chatbots is a language technology application based on AI. We deal with the language diversity in EU by integrating a Machine Translation system into a chatbot platform. We will provide both related work about types of chatbots (e.g. Adamopoulou & Moussiades, 2020), show a demo of a multilingual chatbot developed in the frame of ENRICH4ALL, but also give some food for thought & future ideas for development of chatbots based on big, but also smaller language models, such as one that we created for Luxembourgish. References Adamopoulou, E. & Moussiades, L. (2020). An overview of chatbot technology. In IFIP International Conference on Artificial Intelligence Applications and &Innovations, Springer. Boonstra, L. (2021). Creating a multilingual chatbot. In The Definitive Guide to Conversational AI with Dialogflow and Google Cloud, pp. 187-194. Lommatzsch, A. 2018. A next generation chatbot-framework for the public

Keywords: **Chatbots**, **Public administration**, **Machine Translation**, **Administration**.

International Conference on Innovations for Community Services, Springer, Cham, 127-141. ENRICH4ALL: https://www.enrich4all.eu/ SupWiz: https://www.supwiz.com/