Word Clouds and Beyond: Corpus Linguistic Self-Study Material Package for English for Academic Purposes

Jere Hokkanen

Master's Thesis in English
Spring Term 2019
Department of Language and Communication Studies
University of Jyväskylä

JYVÄSKYLÄN YLIOPISTO

Tiedekunta – Faculty Humanistis-yhteiskuntatieteellinen tiedekunta	Laitos – Department Kieli- ja viestintätieteiden laitos
Tekijä – Author Jere Hokkanen	
Työn nimi – Title Word Clouds and Beyond: Corpus Linguistic Self-Study Material Package for English for Academic Purposes	
Oppiaine – Subject Englanti	Työn laji – Level Pro Gradu
Aika – Month and year Huhtikuu 2019	Sivumäärä – Number of pages

Tiivistelmä – Abstract

Englanti on akateemisen maailman yleiskieli, *lingua franca*. Tässä yhteisössä englannin kieli on välttämätön vaatimus täyteen osallistumiseen kansainvälisissä konteksteissa. Kolmannen asteen oppija tarvitsee akateemista englantia voidakseen kommuikoida: lukeakseen julkaisuja ja kirjoittaakseen kansainväliselle yleisölle. Tässä kommunikaatiotaidossa sanasto on keskeisessä roolissa, sillä akateeminen kielenkäyttö sisältää erikoistunutta sanastoa, ja omanlaisiaan sanastollisia, ja kieliopillisia rakenteita. Lisäksi akateeminen sanasto liittyy akateemiseen ajatteluun: käsitteitä kuvaavien sanojen ymmärtäminen voi olla keskeistä oppilaan oppiaineessa. Korpuslingvistiikan ohjelmalliset sovellukset tarjoavat mahdollisuuden ymmärtää akateemista englantia aidoissa käyttötarkoituksissa, ja toimia apuvälineenä sanaston, käsitteiden ja kirjoittamisen oppimisessa.

Tämä tutkimus, joka on muodoltaan materiaalipaketti, hyödyntää korpuslingvistiikkaa tarjotakseen itseoppimismahdollisuuksia käyttäjille, jotka haluavat oppia akateemista englantia ja ymmärtää oppiaineensa keskeisiä käsitteitä paremmin. Tätä tarkoitusta varten tutkimus tukeutuu myös kognitiiviseen lingvistiikkaan, sillä käsitteiden merkitys mielessä on toinen keskeinen tekijä oppimisen ja ymmärtämisen kannalta.

Asiasanat – Keywords Vocabulary, lexicogrammar, corpus linguistics, cognitive linguistics, academic English

Säilytyspaikka – Depository JYX

Muita tietoja – Additional information

1. Introduction

English has become the international *lingua franca*, the shared common language between L2 speakers, along with the L1 speakers in multinational endeavours such as business and international politics (Rowley-Jolivet 2017: 1). This *lingua franca* position is even more pronounced in the academic community, which is entirely dependent on cooperation across borders (Mauranen 2010: 7). This dependency is particularly true for speakers of medium-sized languages such as Finnish. Finnish language has 5,200,000 native users, (https://www.ethnologue.com/language/fin), which is low compared to relatively large languages apart from English, such as French or German. Therefore, knowledge of this lingua franca is particularly important for an aspiring Finnish academician. This is despite the robust research tradition in Finland in Finnish the more common of the two official L1 languages in the country.

Academic English forms its own distinct own variety of English, *English for Academic Purposes*. It is generally abbreviated EAP. EAP consists of spoken and written communication. The present study, which is in the form of a material package, a portfolio thesis for practical purposes, focuses on the written academic English, or written EAP. The theoretical background comes from the study of corpora: both by *corpus-driven* methods (Biber 2006; Biber and Conrad 2009), and by *corpus-based* method of Systemic Functional Linguistics, SFL (Halliday and Mathiessen 2014). Although academic English, EAP, has existed for a long time, and has a quite strictly delineated structure, it is only modern corpus studies which have identified its distinctive characteristics. These variations include the register or vocabulary employed in Academic texts, and the variations between disciplines (Biber 2006; Hyland and Tse 2007; Hyland 2012). Corpora are also utilized in research and language learning and teaching: the tools utilized in both are similar.

Corpora have been used successfully in the teaching of EAP at the University level, in writing courses where L2 users students compile their own personal corpora. These personal corpora are used for consultation on the linguistic features in EAP (Lee and Swales 2006; Charles 2014). In the previous studies mentioned, the use of corpus tools have been restricted to *WordSmith* and *AntConc*, and these have been utilized to find collocations from a table of concordances. A simple corpus has been used for presentations and introductions, in EAP and outside it: word clouds. Word clouds are based on corpus principles: they turn the most frequent words in a text in a visual representation. They can also be used pedagogically (Filatova 2016), which was the starting point of the present study. However, since word clouds have distinct limitations for linguistic analysis, other tools are included.

For this reason, the present study utilizes a multi-purpose online corpus program, *Voyant Tools*, along with two downloadable corpus programs, *AntConc* and *#Lancsbox*, to utilize word clouds from *Voyant Tools* as a starting point for learner-centred corpus pedagogy and expand from there: to the study of collocations, lexical bundles and to the study of central concepts in each users field of study. For example, a learner, after completing the tasks in the materials included in the appendix can create their own personal corpora of the texts in their discipline, and study both linguistic features and the definitions of central concepts in their field. The learning of EAP vocabulary and the concepts is interlinked: academic language is linked to academic thinking (Nagy and Townsend 2012). Due to this intrinsic link, the cognitive aspects of learning are also covered in the theoretical background, and used as a basis for the materials.

2. Corpus linguistics in English for Academic Purposes

A *corpus* is a collection of naturally occurring texts, or other recorded, stored communication. *Corpora* is the plural form of corpus. 'Naturally occurring' refers to the fact that the texts in a corpus have not been written for the purpose of collecting them into a corpus: compiling a corpus is generally a retroactive process. Typically, a corpus consists of texts that have been selected based on a certain criterion, or certain criteria. As an example of a text corpus, a newspaper corpus includes all the texts from one or more news magazine from a specified time. However, corpora are not necessarily compilations of textual data. Speech corpora are common, and often used to study spoken language variations. Multimodal corpora consisting of videos can be used to study gestures accompanying speech (Coccetta 2011). The present study focuses on textual corpora only, because the focus is on teaching written communication. Corpus linguistics brings statistics and data mining - Big Data methodology - to the study of texts: either for pure quantitative analysis or to support the qualitative analysis. The statistical approach advocated in corpus linguistics is particularly useful in finding generalizations (Brezina 2017), and for providing empirical support to findings about language.

As an example of using quantitative data from authentic texts, Halliday and Mathiessen (2014) studied corpora to uncover grammatical rules from real-life contexts, to create an elaborate classification system of the how grammar of English actually works. Their system in *Functional Grammar* (Halliday and Mathiessen 2014), is capable revealing more regularities out of the system of language than the previous prescriptive methods of grammar ever were – with the added benefit that these categories are proven to exist.

Biber and Conrad (2009) studied corpora to discover what makes academic language separate and distinctive from other language varieties. Discourse analysts and researchers of sociolinguistics study corpora of written or spoken texts to answer their research questions. The present study seeks to utilize and popularize the findings of corpus linguistics, and search for a complimentary link between corpus linguistics and the cognitive factors involving language learning and knowledge building. These findings form the basis of the teaching materials presented in the appendix of this study. The purpose of these materials is to help learners of Academic English to become better at utilizing English for their purposes, and giving them access to tools that build their academic competence.

2.1. Lexis and Grammar in EAP

This study focuses on two areas of Academic English language. First, academic terminology, or *concepts*; and second, *multi-word units*. Together, along with the most frequent words occurring across different academic disciplines (Coxhead 2000), they form the texture of academic writing, the entire tapestry that is a complete academic text. The first category overlaps with knowledge building: the concepts taught in academic studies are *scientific concepts* as opposed to *everyday concepts* based on Vygotsky's division between the two (Swain, Kinnear, and Steinman 2010: 68). The relation between having the vocabulary of concepts and knowledge is reciprocal: to understand a concept, a learner needs to have the prerequisite knowledge and in turn, the understanding of concepts in one's field of study builds the learners knowledge base (Nagy and Townsend 2012: 103-104). The second category, multi-word units, on the other hand, are ubiquitous in academic texts (Biber, Conrad and Cortes 2004; Biber 2006; Biber and Conrad 2009; Weisser 2016). They serve to structure the text, and the thinking presented in a text. Thus, their purpose is more linked to the textual side and thus more purely to corpus linguistics. However, they also serve a purpose in organizing the knowledge in texts, and guide the reader to understand the relationship between the concepts presented in the text.

There are two supporting cognitive frameworks for the learning of a second language, or L2 learning, which are utilized in the present study. Here, they serve complementary purposes, arising from the goals to combine a corpus approach, with the cognitive side of language learning. The cognitive side is necessary for the understanding of concepts without a clear, distinct object to represent them in the physical World. The paradigms chosen for the pedagogical approach here are: first, the *Information Processing* framework, which is in line with computer linguistics utilized here; and second, the *Connectionist* cognitive model (Saville-Troike and Barto 2017: 77), which is in line with the cognitive

aspects of this study. In the framework of Information Processing, IP, learning is learning – there is no separate process for language learning that is distinct from the learning of any other cognitively demanding skill (Saville-Troike and Barto 2017: 77-78). Key ideas in the information processing model are the ideas of *input* and *intake* (Saville-Troike and Barto 2017: 81). Input in learning refers to receiving knowledge and being exposed to it, while intake refers to the process of internalizing the received knowledge. Also, IP stresses that lower-level thought processes need to become automatized for the possibility of higher thought to occur (Saville-Troike and Barto 2017: 78-79). This indicates that for conceptual and abstract learning to occur, basic skill in EAP needs to be at a sufficient level, so that cognitive resources are not constantly divided between understanding the general, non-conceptual vocabulary, and the more complex concepts.

The other approach chosen, connectionism, stresses that language learning is about *strengthening associations* between the various parts in the brain that are involved in the learning of a language. One association is the association between a word and its counterpart in the real world, at least in the case of concrete words (Pinker 2007:9). In the connectionist model, *stimuli* and *responses* are at a key role: repeated responses language stimuli strengthen the links between nodes and units in the brain (Saville-Troike and Barto 2017: 85). This is similar to a model of vocabulary knowledge utilized here. In the learning of vocabulary, Paul Meara's (2007) model of vocabulary knowledge is based on strengths of association and building a strong associational network of vocabulary knowledge. A word activates related words, and the building of vocabulary knowledge is the process of acquiring nodes, the words, and building associations between them. The process, therefore, mirrors that of the connectionist cognitive model closely. One advantage of the two approaches is that neither of the models are exclusive to the learning of a second language, the principles are universal. For this reason, it is not necessary to include a discussion of the roots of L2 pedagogy for the theoretical background section of this thesis. However, some issues related to L2 learning will be discussed in chapter 3 in relation to knowledge building.

In the following chapters, I present brief overview on the role of vocabulary in the learning of *English for Academic Purposes*, abbreviated *EAP* from now on. The overlapping role of *lexis*, or vocabulary, and grammar in a language is covered by introducing the concept of *lexicogrammar*, the combination of lexis and grammar, from corpus linguistics (Biber 2009; Sardinha 2012), and Systemic Functional Linguistics (Halliday and Mathiessen 2014). Then, the role of vocabulary knowledge in EAP will be a topic of analysis. Since it is a central theme of this study, chapter 2.1.3 is not the only place where this idea is explored. Further on, the idea of lexicogrammar is dealt with extensively in the relation it

has with *semantics*, the meaning of words. A hierarchy is built based on the levels of analysis possible, form, meaning, and context. After that, another key theme of the study is explored: the relation of words in texts or corpora have with the concepts and words in the mind.

The next section covers the basic terminology of corpus linguistics, such as *frequency*, *concordance*, *collocation*, and *lexical bundle*. In this section, I review some of the pre-existing literature on corpus-based pedagogy, and present the viewpoint that academic texts can be viewed as networks woven together of concepts and lexical bundles. The idea that texts are 'woven together' is not a novel one: it arises from the very etymology of the word *text*: in Latin, the verb *textere* means 'to weave; to plait together' (https://www.latin-is-simple.com/en/vocabulary/verb/6698/). It is also utilized in Systemic Functional Linguistics in learning how to create texture to writing (Halliday and Mathiessen 2014: 650) Further on, I review some of the indirect uses of corpora: the work done by corpus linguists is already present in language pedagogy, even if corpora are not utilized directly by learners. It is this discrepancy, the fact that corpus linguistics is not utilized more by the learners directly, despite the benefit of doing so, that is at the centre of this study.

After it, in section 2.4., I present how texts can be visualized with corpus tools: what is the advantage of doing so, what sort of tools are available for it, and what are the tools that I have chosen to utilize in the materials. Then, I present a brief overview on existing corpora and review studies where university students have compiled corpora of their own. This knowledge is utilized to guide the users of the materials to further resources or to places where they can find concordances for the words they want to study. Also, it is to underline the benefits of creating a personal corpus for studying concordances from the corpora. The users of the materials practice and create a personal corpus in a task included in the appendix of the present study. However, due to the efficiency of currently existing tools for textual analysis both on single-article and multi-article level, it is not necessary to create a corpus of texts *per se*, a ready-made corpus *for* the users. Rather, the focus is on helping to users create a personalized corpus, following a constructivist principle of knowledge being built. In the present study, there are instructions and tasks that help the users do it by themselves and to utilize these corpora effectively. An additional goal of this study is that if multiple students utilize the tasks presented in the materials, they can also compile a mini-corpus of their gathered writings, whether of their own texts, of others, or combination of the two, for the purposes of textual analysis, so that the users can review their development as writers. In the last subchapter, I review the characteristics of Academic English, its purposes and discipline variations within it.

2.1.1. Vocabulary knowledge in EAP

In order to fully comprehend a text in L2, a reader must know approximately 95%-98% of the words in the text (Nation 2006; cited in Folse 2010: 140; Nation and Webb 2011: 167). This fact, combined with the hegemonic position English has as the Academic *Lingua Franca* of the world (Mauranen 2010: 7-9; Rowley-Jolivet 2017: 1), makes knowledge of English vocabulary practically a prerequisite in any academic learning in L2 (Nagy and Townsend 2012: 92). In the case of understanding a text, this refers to having a *receptive* knowledge of the words (Webb 2008). Even for advanced users of L2, or even L1 users, attaining 95%-98% can be a daunting task. This is especially true when they are dealing with complicated texts with specialized vocabulary, such as academic texts, which contain a high amount of technical, conceptual, and abstract vocabulary (Nagy and Townsend 2012: 92).

In addition to discipline-specific terminology and concepts, academic language contains generalized academic vocabulary, support vocabulary (Schmitt 2010: 79) that one must be aware of and know to be fluent in reading them. Averill Coxhead (2000) in her pioneering work, *The Academic Word List*, created a handy list of this generalized academic vocabulary. For writing, on the other hand, the EAP user needs to have *productive* knowledge of the words. It is an axiomatic truth in linguistics that the productive vocabulary that the user has is smaller than the receptive one – it is easier to recognize a word than to produce it, and far more challenging to *produce* high-quality texts than to read them. Estimates of the difference vary, with earliest studies suggesting that it can be as five times the size of the productive one (Morgan and Oberdeck 1930; cited in Webb 2008: 80). Newer estimates are lower, but the point here is that receptive knowledge is word knowledge is larger than the productive multiple times over.

The most commonly utilized view of vocabulary knowledge, both receptive and productive, includes the amount of words and *word families* known, or the *breadth* of vocabulary knowledge; and the different levels of knowing a word, the *depth* of vocabulary knowledge (Nation and Webb 2011: 226). Breadth refers to the amount of words known, even when the knowledge is only receptive: the user can recognize the word and retrieve its meaning (Nation 2001: 24-25) while depth refers to how many levels of knowledge have about a word: it includes things such as alternative meanings, etymology, and so on.

In order to understand how breadth and depth of vocabulary knowledge are both needed for someone writing in EAP, a metaphor is in order. The writer should be imagined as a boat traversing a river. The river, then, is the vocabulary the writer possesses to navigate the task. Inevitably, the river will not be equally wide and deep in every place that the boat – the writer – has to go to. There are areas that may seem, at the first glance, simple to navigate. There are the areas that may look simple to navigate, places where the river is wide but shallow. However, the boat may get stuck there, as the river lacks the depth for the boat to travel properly in. This is similar to how a writer whose L2 vocabulary knowledge in the topic area is has breath but lacks depth. There are also places where the river is narrow but deep: there, the boat is limited in its ability to move sideways. Instead, it has to travel in a constrained, direct, and limited manner. This is how a writer whose L2 knowledge is deep in their topic area, but lacks general breadth of vocabulary knowledge, has to write: while the said writer may possess great knowledge of the topics and concepts they are experts of, they might lack the breadth to produce texts that are engaging and understandable to read.

A second, alternative view on vocabulary is that of Paul Meara (2009: 76-77), who suggests that instead of breadth and depth, vocabulary size should be measured by *size* and *complexity* of vocabulary knowledge. The size of vocabulary (Meara 2009: 77) roughly corresponds with the breadth of vocabulary knowledge in the first model. However, organisation differs from depth. In the breadth and depth model, the locus of analysis are single words, albeit numerous single words. A high breadth of vocabulary knowledge simply means that there are multiple words known, which are then measured vertically by depth, as in the river metaphor above. In the size and organization model, words exist as a network (Meara 2009: 76) consisting of nodes – the words – which are linked together by associations with related words. The amount of nodes in the network is the size of vocabulary knowledge, while the amount of lines between the nodes refers to the organization. In a well-organized vocabulary, there are numerous lines between the nodes: the language user with an organized vocabulary understands the relation that the words have with each other, and can readily access the words especially if prompted by a related word. This model is particularly well suited for L2 knowledge, as L2 knowledge is not as highly structured as L1 knowledge (Meara 2009: 77; see also chapter 2.1.6. for different models of L2 organization in relation to L1 and the concept that the word deals with). In the following chapter, I review how a text can be similarly viewed as a network: how *lexicogrammar* structures and defines the meaning of the vocabulary in texts.

2.1.2. Lexicogrammar, semantics, and context

In this chapter, I cover three interlinked topics relating to the organization and meaning of texts: lexicogrammar, semantics, and textual context. Lexicogrammar refers to the linguistic structure, where lexis and grammar combine into one (Sardinha 2012: 1). Lexicogrammar can be approached either from the side of grammar or lexis. In this view, grammar is the architecture of a text and words are the building blocks. In the lexis-centred viewpoint, the building blocks – the words – inherently contain information about how they are ordered in relation to each other. In other words, the debate is whether grammar defines lexis or vice versa. *Semantics* refers to the meaning of a word: the study of semantics is concerned with meaning, not form. That is, if such a distinction can be made: form defines the meaning. In this chapter, the relation between lexicogrammar and semantics is explored.

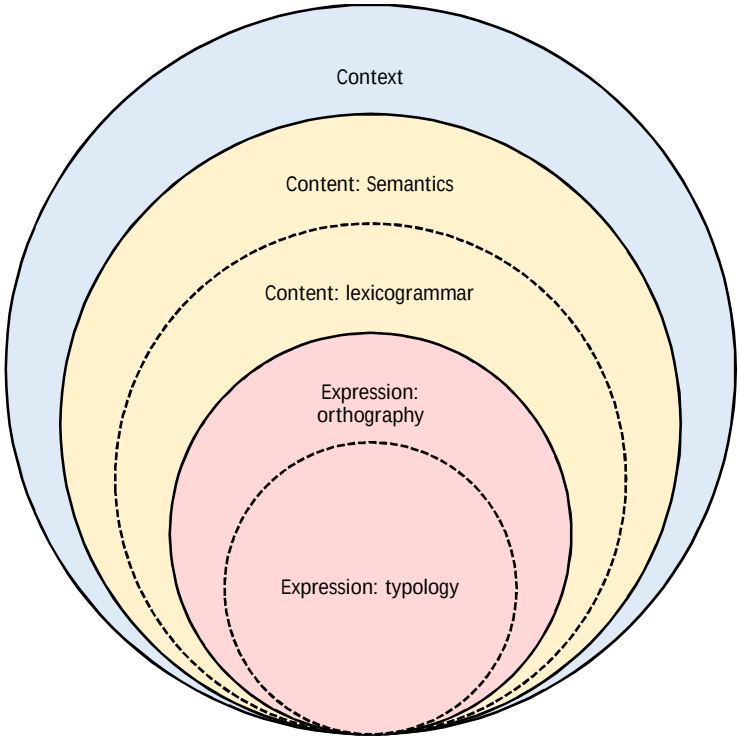
The present study mainly utilized vocabulary-centric approaches to explore academic texts as a genre (Biber 2006: Biber 2009; Coxhead 2012). The tools and classification systems used in Systemic Functional Linguistics were helpful as additional, complementary ways to figure out the effect of form, function, and context in the various genres of academic language. Corpus linguistics has an unfortunate tendency to eschew theory and classifications, and make claims of being solely data-driven (Halliday and Mathiessen 2014: 53). This is not the approach that was chosen here. The claims of corpus linguistics being solely data-driven might be ‘disingenuous and self-deluding’ (ibid.), or reflect a genuine commitment towards openness to new interpretations that linguists who do not specialize in corpora lack (Weisser 2016: 81). The former is certainly not true in the case of Systemic Functional Linguistics, which makes heavy use of corpus data to derive the rules from it (Halliday and Mathiessen 2014: 69-70). In addition, the purpose here is to make the use of corpora accessible and attractive to users who might want to use it something that is only a small part of their research goals if researchers who utilize corpora, or in the case of language learners, as an additional language learning strategy among others. Hence, insights from Functional Grammar (Halliday and Mathiessen 2014) are helpful in providing an antidote against the fragmented nature of corpus studies and provide some hierarchy and coherence to the readers, and the eventual users.

One additional reason for the inclusion of SFL is that Halliday’s work with Systemic Functional Linguistics has contributed much to the teaching and learning of concepts (Swain, Kinnear and Steinman 2010: 68). Ignoring those contributions altogether for the teaching of concepts via corpus linguistics would be a mistake, then. This is because corpus linguistics analyses the product of the writing process (Halliday and Mathiessen 2014: 593), in this case, an academic text. From the

viewpoint of a learner of EAP, gaining an understanding of what an optimal outcome is does not necessarily, by itself, help in understanding the process itself, and how to replicate for one's own needs. Corpus linguistics is best used break down the process of the product, the text, into more smaller, more manageable bits and for analysing it. For the assembling of a coherent product, the process of doing so, SFL seems to be a promising approach. This relationship can be conceptualized as being similar to the inseparable relationship that lexis and grammar have within the sentence level, only in the level of whole discourse and the production of the said discourse, which here is various academic texts.

In making sense of how words acquire meaning in their textual context (Weisser 2016: 207), it is helpful to see lexis and grammar as two ends in a continuum (Halliday and Mathiessen 2014: 56). Although the present study approaches language from a viewpoint that is closer to the pole of lexis than grammar, the conceptual tools developed in a grammatical approach, in Systemic Functional Linguistics (Halliday and Mathiessen 2014) are utilized to explore the complimentary nature of lexis and grammar further. Thus, to illustrate the link between form, meaning, and context, I have adapted Halliday's and Mathiessen's (2014: 26) stratification of *expression* and *content* in speech to written texts into a figure.

Figure 1: Expression, Content, and Context in text. Adapted from Halliday and Mathiessen (2014: 26).



In the figure above, the natural link (Halliday and Mathiessen 2014: 27) that the substrata of typology and orthography have on the one hand, and lexicogrammar and semantics have on the other, is elaborated with a broken line. The larger categories, expression, content, and context, are separated by a continuous one, and shaded and coloured differently. The closest or the most local level categories are in light red, the middle ones in light yellow, while the furthest away, the most global level category – context – is in light blue. In addition, the figure follows the original in also having context as the most global level category. Moving from the local to the global, from particular to the general, or in the figure from smaller circles to the bigger ones, the expression strata consists of two sub-strata: first, graphology in handwriting, or typology in computer-mediated writing; second, orthography. These strata correspond with knowledge of word form in the earlier division of word knowledge, but apply on a broader level. Correspondingly, the content strata expands into two sub-strata: *lexicogrammar* and *semantics* (Halliday and Mathiessen 2014: 25-27).

Contrasted with the typical hierarchy, which is based on the physical organization: word-phrase-clause-sentence-text (Weisser 2016: 160), it may seem contradictory to argue that the lexicogrammatical level is further away from context and closer to the local level of expression than semantics. In systemic functional linguistics, the concise phrase “patterns of wording reflect patterns of meaning” (Halliday and Mathiessen 2014: 27) explains the ordering. This approach is also justified in a lexis-centred approach: from the viewpoint of form and meaning, as in the division between knowledge of word form and word meaning, semantics is meaning-focused in the purest sense, while lexicogrammar is about form, albeit on a level beyond single words. Also, in formulaic language, one has to consider the relation words have with each other, before the meaning and function of the words in lexical bundles becomes apparent (Schmitt 2010: 117-120).

The figure above, moving to the two largest circles, illustrates the corpus-based view of the relationship between semantics and context. The link is not as intrinsic and ‘natural’ as the link between semantics and lexicogrammar (Halliday and Mathiessen 2014: 27), and hence is marked in the figure with a continuous line, but the closeness of semantics and context highlights the principle that words acquire meaning from their textual context (Halliday, Teubert and Cermakova 2004: 38; Weisser 2016: 207). Just as patterns of wording reflect patterns of meaning in the link between lexicogrammar and semantics, words and structures with a particular content reflect the broader context. The context here refers to the fact that certain text types, *genres* (Biber and Conrad 2009: 2), have their own *register* (ibid.), meaning that certain words are characteristic and frequent in certain types of texts (Weisser 2016: 173). In addition, academic language has distinct lexicogrammatical

features, and a distinct way of organizing the texts. These lexicogrammatical features often pose a problem for aspiring academic writers, as they are rather strictly delineated (Biber and Conrad 2009; Charles 2014).

In English for Academic Purposes, understanding and utilizing an academic register is a prerequisite for the learner to become fluent communicator in academic contexts. Further, academic language, content knowledge, and the conceptual tools one needs to think in the nuanced, abstract and precise manner required in academic endeavours, are inseparable (Nagy and Townsend 2012: 93). Therefore, when analysing the meaning of words and structures in EAP, the viewpoint of corpus linguistics, which is text-centric, might not be sufficient alone. It is also necessary to consider how these concepts and meanings exist in the mind, and the interplay between the text and the mind.

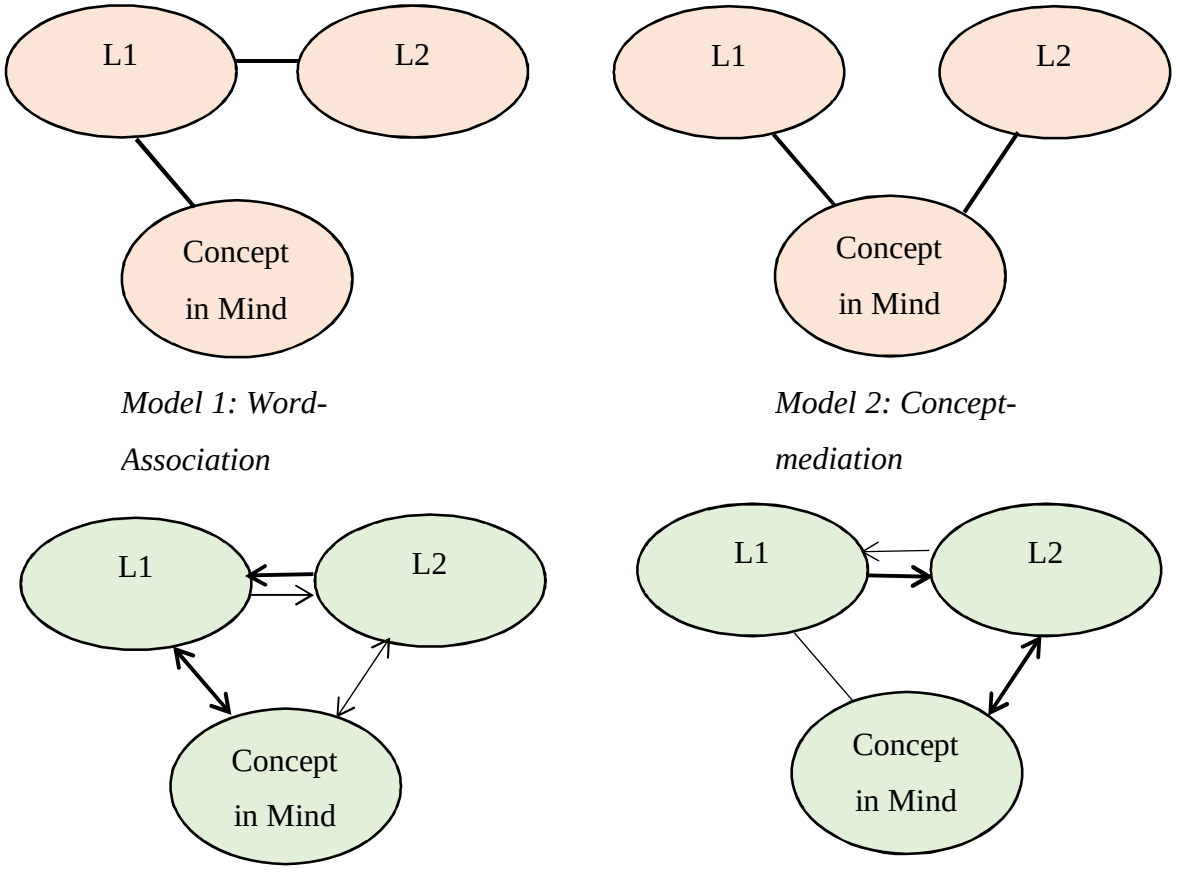
2.1.3. L2 Words across Texts and in the Mind

There is a debate on whether L2 organization in the mind differs from the organization of L1 in the mind (Cook 2002). Earlier, the two main models for organization were the *word-association model* and the *concept-mediation model* (Potter, M.C., So, K.F., Von Eckardt, B. and Feldman, L.B. 1984; cited in De Groot 2002: 37). The same idea was advanced by Weinreich (1953; cited in De Groot 2002: 38) under the name of *subordinate* structure, which corresponds with the word mediation model, and *compound* structure, which corresponds with the concept-mediation model. In word-association model, the L2 word is translated into its L1 equivalent, where it refers to the object or concept that exist in the mind. To use the example of a Finnish native EFL learner, the word 'University' would be translated into the Finnish 'Yliopisto', where it likely conjures up associations about Universities physically, and as places of learning, as places of research and so on.

In the concept-mediation model, continuing with the example of a Finnish native EFL learner, both the word 'University' and 'Yliopisto' would likely bring up the same concept in the mind. The revised model of association, where there is strong association with the concept and L1, and a weak association with the concept and L2 is better in line with newer research (De Groot 2002: 39-41). The relationship between the L1 word and the L2 word, in turn, is strong from the direction of L2 to L1 and weak from L1 to L2. In addition to the three models mentioned, I propose a model for the situation in English for Specific Purposes, of which EAP is a subset, where the link between L2 and concept is stronger, and the relationship between L1 and L2, and L1 and the concept are both weaker or even non-existent.

In my experience as a Master's Thesis student of English, and as an avid reader of English texts from an early age, this situation can easily occur. As an example, for an English learners just learning about *indexicality,* it could be that they learn to understand and utilize the concept before learning the translation that exists for it in Finnish, *aikapaikkaisuus.* Similarly, it is common for Finnish-native computer science professionals to utilize English versions of basic operating platforms, because there is more information available about them in English than there is in Finnish. They can be utilized 'as they are'. This is not to say that the L2 users have become L1 users in these concepts. Rather, it is a case of multi-competence often observed in bilinguals: the mind is constructed differently (Cook 2002: 8). In the figure below, I have adapted the three models described by DeGroot (2002: 37, 39) and created the hypothetical situation.

Figure 2: Four models of relationship of L1, L2, and the concept in mind. Models 1-3 adapted from DeGroot (2002: 37, 39), model 4 hypothesized.



Model 1: Word-Association

Model 2: Concept-mediation

Model 3: Asymmetrical

Model 4: Asymmetrical? ESP/EAP

In the figure, the models that I have chosen to utilize in the present study to conceptualize the relationship between L1 and L2 are shaded green, while the ones that are not used, and are not in line with the newer research are shaded red.

There is evidence that for the learner, the word-association model gives way to the concept-mediation model as the word is encountered and used frequently (De Groot 2002). As a personal anecdote, I have found that some words that I know in L1 and L2, while not working in the word-association model but being instead linked to the concept have minute differences due to the context that they have been encountered. For example, a Finnish adult with interest in planes, who as a child has travelled abroad with the family might associate with the word '*lentokone*' (Finnish for airplane) with a passenger airplane. And the same adult might associate the word 'plane' or 'airplane' in English with a broader class of planes the same person has encountered in books on the topic. This anecdote, while perhaps not sufficient as scientific, empirical evidence to solve the debate one way or the other, serves as an illustration of basic truths from cognitive linguistics: words exist in the broader conceptual networks of the mind, and sometimes in a rather unpredictable way, not as independent dictionary definitions (Pinker 2007: 100-101). It is even possible that the strength, or dominance, of the association in the mind is asymmetric in favour of the association the L2 word has with the concept (De Groot 2002: 38).

The fourth, the hypothetical model I created demonstrates the importance of the context of language use: the L2 words with strong link to the concept have existed in ESP/EAP contexts, where the link to L1 has not been relevant. In addition to the four models presented above, there is a fifth model: a compound model, in which there is not just one conceptual meaning, but multiple, which are linked to both L1 and L2 asymmetrically (De Groot 2002: 49), a model also in line with the fourth one, and the example of the association brought up by the words '*lentokone*' and plane used earlier. It should be noted that none of these conceptual models, not just the one I came up, present a complete account on how work knowledge is stored in memory (ibid.). The last, fifth model, the conceptual, has the advantage of taking into account a recurring a theme that was recurrent in EAP and corpus linguistics: words do not have just one fixed meaning but many.

Corpus linguistics and cognitive linguistics approach semantics from different loci of analysis. In corpus linguistics and corpus-based analysis of language, the locus is on how words acquire meaning in the broader textual context (Halliday, Teubert and Cermakova 2004: 38; Halliday and Mathiessen 2014: 24-27; Weisser 2016: 207). Just as the meaning of words and structures depends on the

lexicogrammatical organization and the broader textual context; words and structures acquire meaning within broader conceptual networks within the mind (Schmitt 2010: 58-62; Nagy and Townsend 2012: 96). Therefore, the complementary nature of corpus linguistics and cognitive linguistics is particularly clear in EAP. For a learner of EAP, implicit or explicit understanding of both types of semantical meanings is essential: the meaning from a broader corpus of studies is essential for comprehension and accurate communication in the learner's field of study (Nagy and Townsend 2012: 91-92), and the cognitive aspect becomes relevant when internalising, understanding, and defining the conceptual meaning in writing. It is not just communication in one's own field, but internalising the knowledge and developing an identity as an expert in one's own field that is at the stake here.

The meaning of a word is often a matter of contention, and being aware of this contention is a reflection of the depth, or the complexity, of vocabulary knowledge. Words are *polysemous*: they have multiple meanings (Schmitt 2010: 52, 54). This is particularly true in Academic English (Nagy and Townsend 2012: 96-97), as the definition of a word, particularly a word conveying an abstract concept varies widely between disciplines; (Schmitt 2010: 77), as all disciplines shape the use of words to suit their own needs (Hyland and Tse 2007: 240). There can even be varieties within disciplines, particularly when a discipline is defined broadly as it often is studies analysing discipline differences (Biber 2006; Hyland and Tse 2007), for example "Humanities". The complementary nature becomes apparent in academic communication: in EAP or ESP, the semantic meaning of vocabulary items in the mind of the user, the cognitive part, should correspond with the semantic meaning the vocabulary items have in the corpus of writing in the relevant field of study, e.g. the commonly used, agreed-upon definition. If it does not, the text cannot fully be considered to be part of the discipline. The variation in semantic meaning inside a single discipline is particularly true within disciplines that deal with topics that are hard to quantify, and encompass many fields of study, e.g. within Humanities and Social Sciences.

Arguably, understanding the various polysemous meanings of central concepts in one's field is what being a professional academic is about. In addition to the multiple meanings a word may have, they carry with them certain connotations from previous usages: from a particular context of use, particular users, or from a particular period in history. Understanding this, and being able to explain the various 'tastes' words have acquired in the *Bakhtinian* sense, that is, the various voices that the word carries, the *polyphony* of words (Mesthrie 2009: 176) is an essential task for anyone aspiring to contribute to academic discourse and become a member of the community (Ivanic 1998; cited in Coxhead 2012:

138). To understand the roots of the concept words and use them in new context is part of the cumulative process of creating academic knowledge. In the words of Isaac Newton, it is to stand on the shoulders of giants.

2.2. Corpus Linguistics and Language Learning

In this section, I overview four basic concepts of corpus linguistics: *frequency, concordance, collocation,* and *lexical bundle*. In addition, I cover the idea of *collocation networks*, which is that texts can be seen as networks of collocations – an idea that is simultaneously present, in a slightly different form, elsewhere in the study. It is present in the view of vocabulary organization in the mind (Meara 2007), in the connectionist cognitive model (Saville-Troike and Barto 2017: 77). It also fits well with the *probabilistic model of knowledge*, which will be covered later. In addition, a brief overview on how corpus linguistics and corpora are and have been utilized indirectly in language learning is presented.

2.2.1. Frequencies

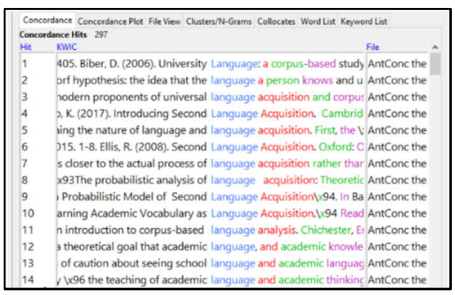
Frequency is a basic, central concept in corpus linguistics (Hyland 2009: 28), which often serves as a starting point for textual analysis utilizing corpus linguistics. In the continuum of lexis to grammar, it is the one most intrinsically linked with lexical knowledge. The analysis of frequency is one of the simplest and quickest tools for analysing a text or a corpus of texts (Baker 2006: 47; Baker 2010: 44). The basic form of frequency analysis looks at how often a word occurs in a text or a corpus of texts, and compares it to the size of the text. Done multiple times, this enables the researcher to identify *key vocabulary* from the text. This method has formed the basis of *word lists*, either for general purpose, academic purposes, or discipline-specific purposes (Nation 2001: 192). A well-known academic word list is The Academic Word List by Averill Coxhead (2000). The latter was compiled by excluding the list of common words in English language from the analysis to identify which words are particularly common in academic contexts. Barring dedicated corpus research, frequency analysis has practical applications for users of corpus tools with different objectives, such as L2 language learners. Online computer corpus applications such as Voyant Tools allow the option for the user to receive information about the frequency of words in a text. This information helps the user identify the topic of a text, or the distribution of topics in a text that covers several.

Frequency analysis is done to identify *key words* in a text, a sample of a text, or a collection of texts (Baker 2010 26-27; Scott and Tribble 2006: 55-56, 74-75). Key words are words that occur in the text most frequently. In an academic text, the most frequently occurring words, not counting articles and conjunctions, are the ones that are relevant to the topic the text covers: hence the term *key* word. The prevalence of frequency analysis in corpus linguistics has led its detractors to label it as simple 'bean counting', not acknowledging the myriad possibilities it offers, such as uncovering the most common language forms for teaching, which surprisingly might otherwise be slighted (Biber and Conrad 2001: 335). However, even frequency analysis, when taken to the next level with modern corpus tools, such as *Voyant Tools*, offers numerous benefits for a researcher or a language learner working with texts, both for reading texts and producing them.

2.2.2. Concordances

A *concordance* is a list of all occurrences of a given *search term*, or *node item*, in a corpus that is typically lined up in a table format (Baker 2006: 71, 76; Schmitt 2010: 123; Baker 2010: 23; Weisser 2016:80). In a table of concordances, the search term appears in the middle surrounded by its immediate context (see Figure 2.2 below), which in this case, simply means words occurring to the left and right of the search term, not the situational context, e.g. from what type of a text is it (Weisser 2016:80). The size of the textual context, e.g. the amount of words surrounding the search term: less than a sentence, a sentence, a paragraph etc. depends on the concordance tool, called a *concordancer* (Schmitt 2008: 123), and the option chosen in it. In the case of a frequently occurring search term, or both, the amount of occurrences can and should be limited, or done in smaller pieces of manageable size for analysis (Baker 2010: 21). This is necessary especially in the case of general corpora containing a huge amount of words. Concordance tables are used for two purposes: getting an impression of the immediate textual context, and finding *collocates* for the key word. Collocates are covered in the next chapter.

Figure 3: Screenshot of an AntConc concordance table of the Word *language*.



The ‘search term’ is often referred to as KWIC, *key word in context* (Baker 2006: 71), as is the case in Wordsmith, or *keyword* in other tools like AntConc. However, to avoid confusion with *keywords* in frequency analysis (Baker 2010: 26-27) and the concept of *keyness* (Baker 2006: 125-128; Scott and Tribble 2006: 55-59; Baker 2010:137-138) in frequency analysis, I use ‘search term’ exclusively for the word a researcher, or a learner, has chosen for concordance analysis. In the following paragraphs, I discuss teaching and learning applications concordance analysis and concordancers offer.

In chapter 2.2.1. I presented frequency analysis as a basic quantitative method in corpus linguistics. For researchers who want to focus on the qualitative side of corpus analysis, *a concordance analysis* can be a useful addition (Baker 2010: 21). Studying concordance tables allows for a more in-depth look on the search term analysed, allowing for qualitative interpretation of the discourses that the search term occurs. In addition, quantitative use of concordance conducted to discover word collocations by an analysis on the strength of association (Schmidt 2010:122-131). While a researcher utilizing corpus linguistics may omit concordance analysis in favour of other tools, in corpus-informed language pedagogy, concordances play a more significant role. This is because concordances are the primary means for learners to retrieve useful data of their target language from a corpus (Ballance 2017: 259; Flowerdew 2015: 110). For example, a language learner wishing to learn the correct preposition to go with a verb (Römer 2011: 215) or before a noun, would use, respectively, a sorting of ‘one to the right’ or ‘one to the left’ of the search term while retrieving data from a relevant corpus in the target language.

As with the use of corpora in general, learners can engage with excerpts of authentic language, produced by native and non-native language users, with concordances. The benefit offered is that learners gain an impression on how language is actually used, which is closer to the actual process of language acquisition rather than learning based on rules and structures (Weisser 2016: 80). In addition, concordances are especially useful for learners who do not have access to native-speaker intuitions about language, as computer corpora can work as a “tireless native-speaker informant(s), with rather greater **potential** knowledge of the language than the average native speaker” (Barnbrook 1996:140; cited in Römer 2011:214; emphasis added). Although learners can access this aforementioned potential indirectly due to the work of corpus researchers, lexicographers, teachers, and material producers utilizing corpora (Römer 2011: 214-217), but there is a benefit in having access to corpus knowledge directly in unpredictable situations that learners encounter in an advanced stage.

Direct use of corpora and concordances seem underutilized in language learning and teaching. A recent research paper on the use of concordances showed that only 2.8% (n=181) of the concordance users used it for language learning purposes (Ballance 2017: 268, 276-277), while 16.0% of the respondents were using concordances indirectly for teaching, and 12.1% indirectly (ibid.). The study used snowball sampling to find users of corpora, starting from known users in a University context. It is possible that the sample consisted of people who are already proficient enough in English, or whose view of themselves as users of language is positive to the degree that they do not perceive themselves as needing corpora specifically for language learning. A study at a different level of proficiency and academic achievement might have yielded different results. Nevertheless, it shows that there is discrepancy between the teachers' own use of concordances, and teaching learners to use corpora (ibid.), and presumably, their own perceptions on the usefulness and willingness to transmit that knowledge to learners.

The discrepancy between teachers' and teacher trainees' willingness to make use of corpus linguistics and concordances directly, a trait observed in previous studies (Breyer 2011; cited in Ballance 2017: 277) and their preference for using more traditional methods in teaching has several explanations. The first one has to do with the interpretation of the results: some of the 12.1% using concordances directly for teaching may have language learners and learner-centred corpus use in mind (Ballance 2017: 271), and the sample of learners, five, is still quite small (ibid.). However, if this is the case, there had not been much retention in the use of corpora by the learners, or the ones who would, at some point, have been in this category have moved on to other categories, such as linguistic or socio-cultural research, or teaching. The second explanation has to do with learner variables, not all learners have motivation, preference or ability to use concordances. Using concordance analysis properly for research can be an arduous, time-consuming task, even for senior researchers (Baker 2010: 21-22). Furthermore, researchers with a background in more theory-driven, structured approaches might be unwilling to adopt a data-driven approach with the constant surprises, regularities, and irregularities that come with concordance analysis (Weisser 2016:81). Language learners likely face similar issues, only in a more pronounced way.

Rod Ellis (2008: 659-671), in his summary of the then-current research on different language learning styles, finds a general distinction between two broad categories of learners. First, *experiential and communicatively oriented*, and second, *norm-oriented and analytical* learner. Although these categories are rather broad and inexact (see chapter 3.3. for more discussion on learner variables), they offer a possible explanation for the low rate of concordance use for learning purposes (Ballance

2017). Comparing this distinction to the results and theories from other sources (Baker 2010; Römer 2011; Flowerdew 2015; Weisser 2016; Ballance 2017), a probable explanation for the low rate of concordance use is that learners in neither of these categories are fully suited and inclined to the type and amount of work required for the use of concordancers. Experiential and communicatively oriented learners might find the opportunities for communication lacking, while norm oriented and analytical learners, like researchers who prefer a more structured approach (Weisser 2016: 81), might find the lack of rules and the inability to find clear binary norms from corpora frustrating.

To make concordances and corpus linguistics, in general, more accessible to learners, an entry point to corpus linguistics should offer opportunities for communication for the first category mentioned, and structure and clarity for the second. The goal should not be to make language learners into dedicated corpus researchers, but to adopt a practical and motivating goal directly linked to their own goals in language learning, and their own contexts of using EAP. Kennedy and Miceli (2017: 111) suggest that the focus for the learner should be shifted from "What can I find out?" to "What can I use?" In their study, successful student users of corpora were motivated, and kept using corpora, once they became aware of its usefulness. This could suggest that in addition to the learner variables mentioned, the lack of students using corpus tools (Ballance 2017) is simply due to the lack of exposure.

In the study by Kennedy and Miceli (2017), a consistent trait with the students that used corpora was that they were using it in a personalized way that was distinct from the way researchers use corpora. Further, the use of corpora was recommended as a method of fostering the language learner's development towards the role of a learner-researcher. (Kennedy and Miceli 2017: 111-112). This suggests that the direct use of corpora and concordance tools can function not only as a native informant for the learner (Römer 2011), but also as a contributing factor in the development of a learner towards a more active, self-directed role. The emphasis on learner-researchers, self-direction, and self-efficacy again demonstrate the mutual link between the corpora and learner, and the importance of these laudable, broader goals in the linguistic environments of the modern society are broadly recognized in modern guidelines for pedagogy, such as the Finnish National Curriculum (OPS 2016).

2.2.3. Collocations

The term *collocation*, and the idea that linguists should examine the "company that words keep", was coined and proposed by John Rupert Firth back in 1957 (McEnery, Xiao and Tono 2006: 82; Brezina, McEnery and Wattam 2015: 139). While there is debate on the exact nature of the term (Lew 2009: 293; Gries 2013: 137; Kang 2018: 86-87), collocation between words means that they occur together with each other more than their individual frequencies would suggest. That is, the two words have a meaningful, statistically significant relationship with each other. (McEnery et al. 2006: 82; Schmitt 2008: 119; Baker 2010: 24; Weisser 2016: 211-212). Due to the ambiguity in the terminology and definitions, Robert Lew (2009: 293) suggests using the term *semantically motivated lexical co-occurrence* for collocations.

In this paper, the use of the term collocation is maintained, but particularly the semantical part of the above definition (Lew 2009: 293) is still useful to keep in mind for the purposes of analysing collocational networks and relationships between concepts: collocations are often specifically studied as *semantic co-occurrences*. This specific semantic focus creates a distinction between collocations, words consisting of two lexemes, and common phrases such as 'of course'. Multi-lexeme words are a relevant issue in formulaic language. Also, prepositions, from the viewpoint of language pedagogy, are an important class of collocations, although in some definitions they do not count as collocates (Brezina, McEnery and Wattam 2015: 140). Modern corpus linguistic software tools, such as GraphColl, use the *Mutual Information Score*, MI score, to identify collocates. In brief, it refers to the likelihood that two chosen words occur with each other. In collocational dictionary work, MI scores are obtained from large generalized corpora.

Not all researchers draw a distinction between two-word collocations, and associative patterns in multi-word units consisting of three or more words, by simply calling the latter *extended collocations* (Hyland 2012: 150). However, while the operating principle behind collocations and formulaic language in general is similar, and much of the theory cited here applies to *lexical bundles* discussed in the next chapter, I opted to use the term collocation for word pairs, with allowances for determiners and conjunctions in some examples. Note that collocation can occur in either direction, and it is not necessary for the words to be immediately adjacent to each other (Gries 2013, Brezina, Enery and Wattam 2015).

For some researchers, there are two schools of thought concerning collocations: first, the *phraseological* one, which is mainly concerned with teaching fluent use of collocations to L2 learners; and second, the *textual* one, which is concerned with empirically examining the co-occurrence of words within texts or corpora (Gries 2013: 138; Nurmukhamedov 2015: 11-12). Although the former approach to collocations was originally for pedagogical purposes and the latter was not, there are several compelling reasons to favour the textual approach. First, the pedagogical framework behind the phraseological approach, as described by Nurmukhadekov (2015: 11-12), was originally from 1940s, when neither computerized corpora analysis, nor easy-to-use corpus interfaces were available. Nowadays, they typically are acquired as a part of the language learning process (Schmitt 2010: 141). Second, advances in computer science, and computerized corpus tools have made the use of the textual approach plausible for pedagogical purposes. Third, the particular type of L2 learning this study is concerned with, English for Academic Purposes, is a highly specialized context of language use, with its own distinctive characteristics that can, and have already been, recognized with textual corpus approaches (Biber, Conrad, and Cortes 2004; Biber 2006; Biber and Barbieri 2007; Hyland 2012; Biber and Gray 2016).

Several criteria for identifying collocations exist. Brezina, McEnery and Wattam (2015: 141-142) present the three traditionally proposed ones to identify collocations, and the three additional ones arising from newer research (Gries 2013). In addition, they propose a seventh criteria. Below, I present a list of the criteria, followed by an explanation on all the individual items in it:

1. distance (traditional),
2. frequency (traditional),
3. exclusivity (traditional),
4. directionality (Gries 2013),
5. dispersion (Gries 2013),
6. type-token distribution (Gries 2013),
7. and connectivity (Brezina et al. 2015).

Distance refers to the span around the *node word* where collocates are sought in. This area, this range of words from the word under analysis is termed the *collocational window* (Brezina et al. 2015: 140). A helpful way to illustrate the collocational window around a node word is the concordance table in figure 2.2: for finding collocations, the node word here is identical to the *key word*. A shorter distance means a stronger collocation (Schmitt 2010: 119-120; Baker 2010: 24-25). Typically, the maximum distance is four or five words from the node word (Kang 2018: 86). For example, working with the

node word ‘coffee’, ‘strong’ is a collocate word that can occur either in the distance of one as in ‘strong coffee’, or within a distance of two, as in ‘*strong* black *coffee’*.

Frequency refers to how typical the association between the words is (Brezina et al. 2015: 140), e.g. the frequency of the word pair in relevant corpora. For example, ‘strong’ and ‘coffee’ occur frequently with each other (Online OXFORD Collocation Dictionary). The existence and the frequency of use of common collocations in English likely explains the hypothesized *lexical priming* in collocates for native speakers (Durrant and Doherty 2010: 127-129). One focus area for the *phraseological* L2 pedagogy is teaching frequent collocations that natives would use (Nurmukhamedov 2015: 11-12).

Exclusivity means that the words occur specifically together. For example, prepositions generally do not count as collocates with the noun they occur with, such as ‘in love’, because *in* is a frequent word elsewhere (Brezina et al. 2015: 140). Of course, preposition and noun pairs are still an important feature of formulaic language and L2 learning, as mentioned in chapter 2.2.2.

Directionality of collocation is about which word in the pair occurs before, and which after (ibid.). Gries (2013: 141) argues that the traditional association-based measurement of collocation conflates bidirectional/symmetric probability of association with the unidirectional/asymmetric association. Bidirectional means that the word can occur either before or after, while unidirectional means that it can only occur either before or after. Therefore, directionality of collocation should factor in the analysis, and there are methods to measure it (Schmitt 2010: 131-132; Gries 2013). Collocates often have a certain order in which they appear, like ‘strong coffee’ rather than ‘coffee strong’. Directionality is a particularly strong feature in idiomatic collocations, or collocations with both a literal and idiomatic meaning (Macis and Schmitt 2017: 324), For example, ‘salt and pepper’, rather than ‘pepper and salt’. This collocational direction occurs whether the meaning is idiomatic or literal, in the example, either when referring to the spices or the colour of someone’s hair or facial hair.

Dispersion refers to how the node word and the collocate are dispersed in the corpus, meaning how many cases of the collocate are occurring, and in how many texts they occur (Brezina et al 2015: 141). The advantage of this measurement is that it tells whether the collocation is of general nature or specific to certain texts, where it occurs more frequently than in others.

Type-token distribution of collocation measures the strength of collocation and the competition that a chosen collocate faces with other possible types of collocates for ‘the slot’ next, or to a specified

distance, from the node word (Brezina et al. 141-142). For some node words, there are multiple possible collocates that can be linked to the word. For example, using the node word coffee, the slot before it could be filled with ‘strong’ or ‘hot’, or even ‘Irish’, when referring to the specific drink.

Connectivity expands the two-word criteria presented above, by introducing *collocational networks* (Brezina et al. 2015: 142). Collocates do not occur in isolation. Instead, they occur in the broader context of the text. Further, collocates form chains, creating semantic networks in the text (ibid.). This collocational measure, the idea that texts consist of semantic networks of collocations (Philips 1983, cited In Brezina et al. 2015: 139) is particularly important in this study and has a chapter devoted to it, 2.2.5. As with concordances, the focus here is in the practically-oriented dictum “what can I use” (Kennedy and Micali 2017: 111). The visualisation of text to a semantic network is explored later in chapter 2.3.2.

A detailed explanation on how collocations are calculated is beyond the scope of this study. The concept of *strength of association* (Schmitt 2010: 124-131) is central in collocations. However, as the newest research on collocations identified seven criteria for collocations, and the term by itself is controversial (Brezina et al. 2015; Kang 2018), it is not necessary to explicate the mathematical formulae behind the measures here, as the mathematical side is nowadays handled by computer programs. The focus is on the role that collocations play in academic discourse, and on raising awareness about collocations, serial collocations and collocational networks, and larger units of *formulaic language*.

2.2.4. Lexical Bundles and Phraseological Units

Formulaic language, in addition to the collocations discussed in the previous chapter, consists of frequently co-occurring words that number more than two. For example, there are frequently used, important and ubiquitous expressions in everyday communication such as “How are you?” and the many, formulaic counterparts it has, such as “I’m fine, how are you?” that form a full sentences by themselves. They serve an important function not only in communication, but also in the acquisition of a language: language learners initially acquire language in chunks of formulaic language that serve a communicative function (Schmitt 2010: 139-140). Correspondingly, learning to use these phrases form an important part of L2 pedagogy especially in the early stages. In EAP, a harder skill to master is the use of appropriate formulaic expressions within sentences. To identify key characteristics when it comes to frequently occurring word combinations within a particular register, corpus studies have

identified chunks of words that frequently occur together (Biber et al. 2004). These chunks serve important functions in a text such as organizing it, and are characteristic of all texts that consist of full sentences, paragraphs and so on, such as essays.

According to Martin Weisser (2016: 162), multi-word units consist of three categories. Prefabricated expressions, which is also includes the full sentence level expressions, *lexical bundles* (Biber, Conrad and Cortes 2004; Biber and Barbieri 2007; Biber 2010; Chen and Baker 2010), and *phraseological units.* All these units serve an important functions in discourse, such as referring to other parts, clarifying, or presenting. The difference between the latter two is that phraseological units conform with grammatical constructions (Ädel and Erman 2012: 82), such as with "to be more precise" whereas lexical bundles necessarily do not (ibid.). Thus, lexical bundle is a more inclusive term: it can include phraseological units. Or, to conceptualize it with the help of Systemic Functional linguistics, it can be said that both phraseological units and lexical bundles are *logico-semantical* ways of providing cohesion to text. For example, "to be more precise", a phraseological unit in the system of conjunction, can be classified as within SFL as being *conjuncted, elaborating, clarifying,* and *corrective* expression (Halliday and Mathiessen 2014: 612). Lexical bundles, on the other hand, are less grammatical and instead provide lexical cohesion in the text (Halliday and Mathiessen 2014: 642-644), in addition to serving functions such as elaborating and so on. Also, there are lexical bundles that are simply technical terms consisting of many words, which generally are not considered to be lexical bundles in the sense that corpus linguistics classifies and studies lexical bundles (Biber et al. 2004: 372-373).

Of all the terms presented, the term lexical bundles is the one chosen for this study, because they are recognizable parts of a particular register (Biber 2006; Biber and Barbieri 2007; Biber and Conrad 2009). Here, the focus is on helping the users of the materials presented acquire both conceptual knowledge and to utilize lexical bundles and lexicogrammar to tie the writing and thinking together. Lexical bundles are frequently occurring series of words that occur in a given text or corpus. One important distinction is the distinction between two broad categories of lexical bundles: the ones conveying content, which include idioms and multi-lexeme terms, and functional lexical bundles that have a purpose related to the organization of the text, for example common expressions in academic writing such "this enables us to", or "as a consequence of". Here, the approach is the exact opposite of the one regarding collocations. With collocations, the focus is on *semantical co-occurrence*. With lexical bundles, the focus is more on *lexicogrammatical co-occurrence*.

Regardless of the distinction, multi-word units consisting of more than two words have their place in EAP. In technical fields, they are obviously of particular importance. However, research indicates that the learning of technical terminology might be easier for learners (Schmitt 2010: 77-78). Nevertheless, for the novice academic, adopting field-appropriate idioms is a step in becoming accepted in the scientific community and in developing an identity as a researcher (Ädel and Erman 2012; Nagy and Townsend 2012). This is similar to language learning in other contexts, as the process of acquiring formulaic sequences is an aspect of language learning that requires sociocultural adaptation (Dörnyei, Durow and Zahran 2005; cited in Schmitt 2010:140). Learners who lack an experience of sociocultural adaptation may find this aspect of EAP particularly challenging.

In addition to the distinction between idiomatic bundles and the functional ones, functional lexical bundles can be further separated into categories by their purpose, or predominant function in a given discourse. The classification systems vary between corpus researchers (see for example Biber, Conrad and Cortes 2004; Cortes 2004; Simpson-Vlach and Ellis 2010; Ädel and Erman 2012; Halliday and Mathiessen 2014: 652). The classification system presented below is the one adopted by Biber, Conrad, and Cortes (2004):

1. Stance expressions,
2. Discourse organizers, and
3. Referential expressions.

The use of certain lexical bundles, and formulaic language in general, is ubiquitous both in written and spoken EAP, although they are slightly less in academic prose and textbooks (Biber and Barbieri 2007: 273). Of course, the types of bundles are different depending on whether it is in spoken or written communication, and on the discipline. For example, academic teaching in humanities and social sciences favour hedging, and lexical bundles related to hedging, to acculturate students to understand the non-binary relationships between various concepts (Hyland 2009: 101).

In chapter 2.1.3., I analysed how the semantic meaning of the word in the cognition of the learner should correspond with the meaning or meanings it in the relevant context, and how developing this relationship is at the heart of attaining academic knowledge. Similarly, lexical bundles highlight key information from a text, and construe the written or spoken academic discourse in a way that encourages the learners, and the readers or listeners, not just to consider facts, but also to think about the complex relationships between the topics (Hyland 2009: 102). If academic texts consist of networks of words that collocate with each other (Phillips 1983; cited in Brezina et al. 2015: 139),

then functional lexical bundles serve to structure this collocational network, which is the topic of the next chapter. Similarly, this structuring helps learners to structure the argument of the text in their mind, which is the textual equivalent of a mind map. So, if the single words and multi-lexeme technical terms serve as the 'nodes' in the network, then the functional lexical bundles serve as the 'lines' that link the concepts together.

The notion mentioned above gains theoretical support from the notion of *lexical chains* in Systemic Functional Linguistics (Halliday and Mathiessen 2014: 652-654), which conceptualizes frequently co-occurring collections of words thusly. It also gains empirical support from studies where groups of users are compared on the basis of their use of lexical bundles. Chen and Baker (2010) compared of the types of lexical bundles in academic discourse native experts in a field use compared to students on the same field of studies. They found out that students overuse discourse organizers, and underuse of referential language compared to the experts. Support for the findings of Chen and Baker (2010) was found in a similar study by Ädel and Erman (2012), comparing the types of lexical bundles used by English-native and Swedish L1 non-native higher education student writers. The general pattern was same, with native writers using a wider variety of lexical bundles than non-native user did (Ädel and Erman 2012: 87). However, the discrepancy between the type and amount of lexical bundles was greater, and there were exclusively native and exclusively non-native bundles, in addition to the shared ones (Ädel and Erman 2012: 89). In addition, the non-native students, once again, used discourse organizers more.

Hence, I hypothesize that the observed tendency of students to overuse referential lexical bundles in writing is due to two reasons. The first is due to insufficient specific language skills, or a transfer of language skills from other contexts of English use that are not suitable in EAP. Second, it could be an attempt to organize the network of concepts in the mind, similar to how coherent texts consist of collocational networks (Anderson 1983; cited in Brezina et al. 2015: 139). To demonstrate this idea, I return to the idea of concepts as 'nodes', and the lexical bundles as 'lines' in the structure of the text, similar to a mind map. As the concepts, the 'nodes' they are dealing with, are yet an unfamiliar territory, student writers are uncertain of how to utilize the 'lines' to link the concepts in their discourse together and to explicate the position of concepts to each other. Also for themselves. Consequently, there are more discourse organizers.

Currently, there appears to be a lack of empirical studies exploring the *reasons* for this observed discrepancy beyond a simple lack of practice, although it is well documented (Hyland 2012; Ädel

and Erman 2012). It seems that there is simply an assumption that during the course of writing assignments and involvement with academic texts, aspiring users of EAP will implicitly adopt the norms regarding the use of lexical bundles, e.g. start using referential bundles and eschew discourse organizers, along with other forms of sociocultural adaptation to the norms of the academic community. The types of lexical bundles used by professionals, referential lexical bundles such as "in the context of", or "as shown in the figure" do not exist simply for stylistic reasons, but are part of the argument of an article or other academic writing. In other words, they are part of the critical thinking required in academic endeavours (Nagy and Townsend 2012: 92). They also signal competent participation in the academic community, as they are so natural in academic discourse (Hyland 2012: 165).

An alternative, or at least a complementary, explicit EAP-learning centred approach would recognize the importance of attention to professional lexical bundles. The finding that students sometimes use lexical bundles that professionals never use, and vice versa, students might not employ bundles frequent used in academic discourse at all (Cortes 2004; cited in Hyland 2012: 159), combined with the finding that technical terminology is learned the easiest (Schmitt 2010: 79) suggests that explicit attention to lexical bundles is needed. However, as language is so intrinsically linked with cognition, and in the case of EAP, abstract, high-order thinking; these lexical bundles should be encountered in authentic contexts, not solely as decontextualized lists containing the most frequent lexical bundles. This is particularly important due to the ubiquity of referential lexical bundles in academic writing (Hyland 2012: 158-159). Learners should not only be exposed to what lexical bundles in EAP are, but what functions they fulfil in the text: how professional writers use them to construct an argument and to tie in the concepts together in a fashion that is clear, and to the reader, seemingly effortless.

2.2.5. Collocational Networks

As alluded earlier in chapter 2.2.3., which was about two-word collocations, collocations can form *collocational networks*. These networks allow for analysis of something beyond collocation pairs and more similar to lexical bundles, that is, word *troops* (Baker 2016: 140). Compared to lexical bundles, word troops vary widely between texts, even in the same genre. The troops of words formed by collocational networks describe the content and the topic of the text, being more related to the *concepts* explored in an article, rather than serving a functional purpose like lexical bundles do. Not like lexical bundles, which function as stance expressions, discourse organizers, and referential expressions (Biber et al. 2004). To return to the metaphor of the mind map presented earlier in 2.2.3.,

where the nodes are concepts and the lines are the supplementary discourse, word troops displayed by a collocational network give a fair impression of what the nodes in the map are like.

Collocational networks are relevant both in research and pedagogy, or both, as the distinction between the two can be blurred particularly in the case of EAP. In research, collocational networks allow for an analysis of the interlinked discourses in a text, such as newspaper articles about a certain topic. For example, Paul Baker (2016) in his introductory paper on the use of AntConc and GraphColl for the analysis of collocational networks described earlier in 1983 by Sinclair (Brezina et al. 2015: 139), displayed his work on collocational networks in analysing newspaper articles from the conservative British tabloid newspaper Sun about Muslims (Baker 2016:142-159). He had studied the collocational networks formed by the word *troops* in the articles, not to be confused with the earlier-mentioned concept of *word troops*. While utilizing GraphColl, he started with the node word *troops.* By clicking on *troops*, the network showed the collocates for troops: *British*, *our*, and *Afghanistan*. GraphColl was then used to display the collocates for the latter, three nodes further, expanding the network.

The advantage of GraphColl for analysing texts is that it illustrates multiple possible collocates at once. For dedicated corpus research and the analysis of strengths of association, mutual information score, etc. it is often be supplemented by another tool. For example, in the abovementioned example, Paul Baker (2016) used a popular freeware corpus concordancer tool AntConc for the precise analysis of specific terms, with GraphColl serving in the function of providing a more global-level starting point for the analysis, and for illustrating the linked nature of the terms under analysis. This order can also be reversed, with a concordance table used first, and GraphColl second, to see how the term under analysis forms links and mutual collocates in the text.

GraphColl does offer a possibility for the user to obtain MI scores, but as noted by Baker (2016: 146), as a tool, it did not fulfil the criteria of *directionality* (Gries 2013). Directionality is presented as the fourth measure of collocation here in chapter 2.2.3., on collocations, as first of the modern measures of collocation identified by Gries (2013). Therefore, GraphColl does have its limitations that should be kept in mind when using it apart from the other tools. However, for computerized corpus pedagogy, GraphColl and collocational networks are particularly useful in exploring phenomena that are linked together. Also, nowadays, GraphColl is part of a larger corpus toolkit called #LancsBox, which offers more additional tools. GraphColl is particularly useful in analysing ideologically loaded words, where the words under consideration would serve as the starting point nodes for the analysis, in a manner similar to how Baker started with *troops* and expanded further and further. It also has potential

in concept-rich fields of academic study. In those fields, the nuances in the definition of the concepts are a central part of understanding the topic and the field. In those fields GraphColl has particularly strong potential for the study of concepts in Academic English.

It is perhaps relevant to note that GraphColl was originally created as a corpus tool for Social Sciences at Lancaster University. Social Sciences certainly meet the criteria for being context-rich and nuanced in definitions. Of course, the distinction between learning academic language and learning the academic discipline itself, and learning to think in the required manner is once again practically non-existent (Nagy and Townsend 2012). In this material package, the concept of collocational networks and the tool *GraphColl* is explored for in-depth content learning. Thus, while lexical bundles and collocational networks are both tools for analysing and subsequently learning to use multiple words that are linked together in one manner or the other, collocational networks – once again utilizing the distinction in Systemic Functional Linguistics – are clearly connected more to the *semantics* of a particular field of EAP, while lexical bundles deal with the lexicogrammar of EAP. Or, in a broader sense, collocational networks are clearly closer to the pole of lexical analysis while lexical bundles demonstrate an aspect of EAP grammar. More detailed commentary on the use of GraphColl, and an analysis of its technical features, is presented in chapter 4. In addition to GraphColl, Voyant Tools is capable of displaying collocational networks with the Links option.

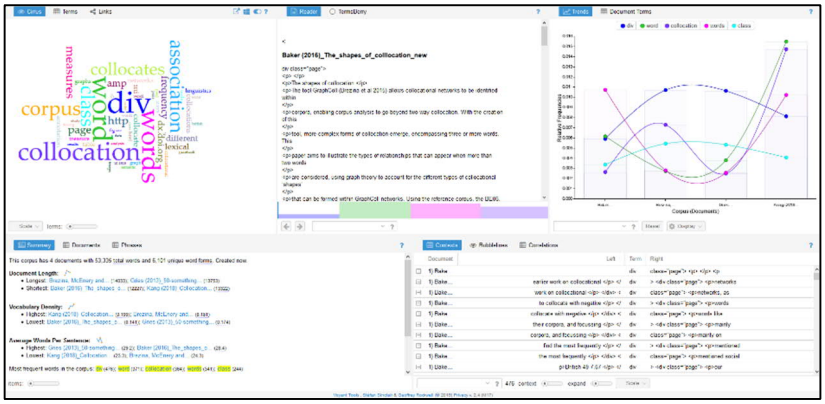
2.2.6. Indirect use of corpora

In addition to research and language teaching and learning, there are other common uses for corpora that indirectly relate to the learning of language both in L1 and L2. For example, corpus work has been used to identify the most common academic words generally (Coxhead 2000; 2010), which have been used in teaching vocabulary. They have been utilized by lexicographers in their work to compile native-language or L2 dictionaries, especially for dictionaries that contain examples of using the word in real-life situations. This background work, in turn, is utilized in language teaching and learning. In addition to having dictionaries and thus utilizing corpora indirectly, the work of corpus linguists can have an indirect effect on language teaching through the use of materials writers (Römer 2011: 207). In the case of using corpus linguistics as a basis of the materials, the examples come from real-life, authentic situations, instead of situations that are generated solely for pedagogic purposes. The utilization of corpora can then have an effect on language syllabi as well as reference materials and the like (ibid.).

2.3. Making text analysis more visual with corpus tools

In this chapter, I review the use of corpus tools from the viewpoint of utilising them as a strategy for studying and researching texts. Mainly, that they can be used to complement the reading of texts by returning data of the texts that can be used either to focus on key areas inside the text, or present an overview of the linguistic features in them. In addition, the use of corpus tools adds another modality to the analysis: the representations created by it are visually engaging, as seen in Figure 4 below, a screenshot of Voyant Tools page analysing a mini-corpus.

Figure 4: Screenshot of Voyant Tools page after texts have been uploaded.



Visualisation of text is a common occurrence in presentations. Often, a *Word Cloud* is shown before the presentation in order to orient the listeners to the topic at hand. Word cloud generates a visual, often coloured, representation of the common terms in a text, where the more frequently occurring words are larger, and are in a more central position. A typical word cloud shows words both horizontally and vertically, but there are word tools which allow for the option of diagonally positioned words. In addition to word clouds, there are other text visualisation tools utilizing corpus linguistic principles, such the trends chart and TermsBerry in Voyant Tools, and GraphColl. In this portfolio thesis, I explore the possibility to use text visualisation tools to teach frequencies, collocations and collocational networks, and lexical bundles. While concordance tables are not particularly visual, they are still an efficient way of making collocations to a word visible and separating them from larger chunks of text. Further, AntConc does add colour to clarify collocates, and allows for the analysis of several collocates at once, all marked with a different colour. This helpful addition makes the task of looking at a concordance table less strenuous. Wu, Provan, Wei, Liu, and Ma (2011: 742) divide the existing word visualisation methods into three categories of orientation:

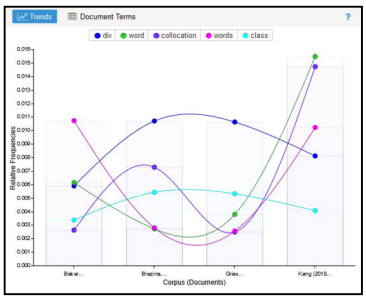
1. Word-oriented,

2. Document-oriented,
3. and theme-oriented methods.

Word-oriented visualisation method generate a visualisation, typically a word cloud to summarize the document, but often do not offer any additional information about the relationships that the words have with each other, or the temporal location within the text, e.g. whether they appear in beginning, in the middle, or at the end of the text (ibid.). The word clouds generated by the tool used in the present study, Voyant Tools, do not provide this semantic and temporal information about the text. However, Voyant Tools provides a word-oriented visualisation tool that tracks the temporal location of the words in the in the form of a graph, which tracks where the chosen words occur within a text. With this, the user can select the words central in the word cloud for temporal analysis. In the graph, the horizontal x-axis tracks the temporal position in the text, and the vertical y-axis indicates the variety unique vocabulary in the part of the text.

Therefore, the inclusion of a more sophisticated form of creating word clouds is unnecessary in the present study: the temporal information can be acquired elsewhere. In the general overview, using the graph, Voyant Tools characterizes a high amount of unique vocabulary items as “professionalism” of the text. This definition may be problematic in certain text genres, such as academic thesis work, as there may be a high amount of repetition of vocabulary items due to the need to explain the meaning of these items. However, in learning productive vocabulary and writing, the tool does allow the user to notice any discrepancies in the text they produce and make the end product consistent in the utilization of register, and may inspire the user to pay attention to their style more.

Figure 5: Screenshot of Voyant Tools trends graph.



Document-oriented visualisation methods transform a collection, or a corpus of texts, into a visual representation. This allows the user to easily compare the register and style of multiple documents. For example, when using TermsBerry, an option on Voyant Tools, the user can click on the words

and see the amount of documents it appears in. When comparing documents in the same genre or a subgenre, TermsBerry allows the user to gain an understanding of the register in the genre or subgenre. It is likely that words characteristic of the genre appear in multiple documents. With *TermsBerry*, there is an added benefit: when analysing academic documents about the same topic area, the user will have the opportunity to identify the concepts that frequently appear in the papers. This last advantage conferred by TermsBerry overlaps with the third and last category. The trends chart presented in Figure 5 also allows document-oriented analysis: the screenshot shows the frequencies of the five most common words in the four articles of the mini-corpus. There are other options in Voyant Tools that allow for document-oriented visualisation methods, which are explored in the appendix in detail.

Finally, *theme-oriented* visualisation methods derive information about the semantic content of the document, and then present the results in a visual form (ibid.). In addition to the tool mentioned by Wu et al. (2011: 742), *ThemeRiver*, and the one they propose: semantically linked word clouds based on *Seam Carving*, Graphcoll counts as one. The difference from GraphColl to the ones mentioned is the level of analysis: while ThemeRiver and Semantic Carving Word Clouds analyse texts in a top-down manner, or global level of analysis, GraphColl uncovers themes on a local level, starting from the local level of single words. Nevertheless, I've opted to include it in the categorization, as the end result of a GraphColl map of collocations is rather visual. It functions much better as a theme-oriented analysis tool as a word-oriented one, and since TermsBerry, which doubles as document-oriented analysis tool is currently the only one used for theme analysis, the inclusion of a second tool with a different focus is justified. This process is explained in more detail in chapter 2.2.5. I have created a table of the categories and the tools used for their analysis, complete with the loci of analysis, examples, and the ones used in the present study in a table format below.

Table 1: Text visualisation methods, their loci of analysis, examples of tools and the ones used in the present study.

Category:	**Typical locus of analysis:**	**Examples:**	**In the present study:**
Word-oriented: frequency	Single text or a corpus of texts	Word cloud in *Voyant Tools*, *Wordle*, *WorditOut*, *ManiWordle*	*Voyant Tools* word cloud

Word-oriented: collocation	Single texts or a corpus of texts	*AntConc*, *GraphColl* and *Voyant Tools* options	*AntConc*, and *Voyant Tools* options
Document-oriented	Multiple texts	*TermsBerry* and other options in *Voyant Tools*, *FacetAtlas*, *Seam Carving* word clouds (Wu et al. 2011)	*Voyant Tools*
Theme-oriented (adapted version from Wu et al. 2011)	Semantic relationships, global or local	*Themeriver, Seam Carving* word clouds (Wu et al. 2011)	*Voyant Tools*

2.3.1. Making frequencies more visual: word clouds and trend chart on Voyant Tools

Word clouds, such as *Wordle*, *WorditOut* or *Voyant*, are free online tools that visualize the vocabulary of a text. They use simple data mining to find the highest frequency words in a text, and then create a visual representation of the text by placing the words in the form of 'a cloud' of words. More frequently occurring word are in a more central position than the more seldom occurring words. As the main ideas of a text are likely to appear several times in a text, word clouds offer a rough summary of the text, and present a tool for introducing the ideas of a text to the audience before reading the text. Further, the user of a word cloud tool can customize the cloud by removing words from the cloud and deciding on the arrangement and colors. They are often used in posters, presentation slides, and in websites or blogs to present the main ideas of the text. This is done to prime the audience for the main terms covered in the presentation. They can also be utilized in language learning and teaching: both in teacher-led activities and as learning tools for the students themselves. The present study focuses mainly on the latter.

Figure 6: Word cloud of the present study made with Voyant Tools.

Depending on the program used, the user of a word cloud tool has the option to customize certain aspects of the cloud. Unfortunately, the tool used here, Voyant Tools, only allows the option of choosing the amount of words. The amount from 25 to 500, obviously limited by the length of the text and the amount of unique word forms in it. While Voyant Tools may seem lacking in this aspect, it more than makes up for this with various other frequency analysis tools, such as a list of the most common words in a document and TermsBerry, which will be discussed in the next chapter, among others. While the word cloud option from Voyant Tools is the word cloud tool chosen for the tasks in this material package, some comments on the general possibilities offered by other word cloud tools are in order. An interesting concept, but unfortunately beyond the scope of the present study, are *semantic preserving word clouds* (Wu, Provan, Wei, Liu, and Ma 2011). They allow for comparative analysis of documents, and sort the cloud by topic areas by arranging semantically similar words close to each other (Wu et al. 2011: 741-742). This option of thematic analysis in *Voyant Tools* is handled via other tools, not the word cloud. However, *Voyant Tools* is designed to function as a package: by changing the parameters on one of the tools provided the user also influences the view it offers with the others.

Olga Filatova (2016) discusses using Word Clouds as a teaching tool in ESL instruction for all levels of proficiency, including university, to improve students' reading and writing skills. For vocabulary instruction, students can utilize word clouds to improve both their receptive or productive vocabulary skills. Receptive vocabulary is particularly important in reading comprehension (Webb 2008), whereas a broader productive vocabulary makes the writing more fluent. The instructional strategies she proposes can be classified as explicit learning strategies, but they can be integrated into the students' interests and studies, such as reading articles in English on their field(s) of study. The basic model she proposes can be summed up as follows (Filatova 2016: 440-441):

1. An instructor, or the student, generates a word cloud of an article to be read
2. Students discuss the text
3. Students looks up the unknown words from the cloud
4. Students read the article with the added benefit knowing the word beforehand
5. Contextual comprehension is increased and reading time is reduced
6. Students can better outline and summarize the text, or formulate a response to it.

Here, the basic task is within the context of language instruction with the added benefit of social interaction and peer support. Of course, there is no reason why students could not do it by themselves

or adopt the use of word clouds as a strategy, as experimenting with the use of word clouds is encouraged (Filatova 2016: 447-448).

The pre-reading word cloud task finds anecdotal, but professional support from Olga Filatova's experience as an ESL instructor (Filatova 2016: 440, 448), and from research on cognitive processes involved in learning new vocabulary items. The strategy of creating a word cloud and looking up the unfamiliar words from there is an explicit strategy, but like other explicit strategies (Ender 2016: 538), can promote incidental as well intentional learning of new lexical items. The word cloud obviously makes the recurring word salient, thus combining the two factors in Rieder's (2002; cited in Ender 2016: 538) three criteria for a new vocabulary item to become in the focus of attention. Presumably the third one, content-based necessity, is also made more salient by the word cloud (Filatova 2016: 440). This is especially when the articles are in the students' field of studies, and the student also has the expertise to assess the necessity of the item.

There is a benefit that the word cloud, and consulting the dictionary on the words in the cloud over explicit vocabulary strategies to be used while reading, such as consulting the dictionary when one encounters an unfamiliar word. It may alleviate the "beginners paradox" of a student not reading texts due to being discouraged by having insufficient receptive vocabulary knowledge, and thus failing to get enough input, and failing to build their vocabulary by utilizing the implicit strategy of extensive reading to learn words (Hunt and Beglar 2005: 27). Also, in the case of English being an absolute requirement in the student's field of study, this vicious circle extends to content knowledge. Although at the level of tertiary education, students can at least consult paper or online dictionaries, it is possible that knowing beforehand that one has to take the time to do so may still discourage the learners from the reading task. In contrast, the pre-reading word cloud proposed by Filatova (2016: 440) may reduce reading time, as the students become familiar with the unknown words beforehand. As ignoring an unfamiliar word is a frequently used strategy – naturally so when reading is not done for the sole purpose of acquiring vocabulary – the pre-reading word cloud guides the reader to focus on what is relevant, and to ignore the more peripheral vocabulary.

For learners of different types, ages, and proficiency levels, notwithstanding the question whether the theory of different learner types, visual, auditory and kinesthetic is correct, Filatova suggests that word clouds may be particularly appreciated by visual or kinesthetic learners, as well as younger students (2016: 446). They may be beneficial to students of all proficiency levels. This explicit strategy may be particularly useful for lower proficiency students, or students struggling with

motivation. ESL or EFL readers often use inferring from the context as an effective strategy for retention (Ender 2016). Inferring does have its limitations, it is not particularly efficient strategy for students of lower proficiency (Hunt and Beglar 2005: 27), and inferring is solely focused on meaning, not form, and therefore might not leave a strong trace in memory (Ender 2016: 555). It might fail in building associations in vocabulary knowledge and be insufficient in building up the learners' vocabulary: the learner is merely benefitting from the fact that they understand 95-98% of the vocabulary in the article (Folse 2010: 140). Although understanding how words relate to their context is of paramount importance, a decontextualized lexis has benefits that extend to other contexts. The word cloud helps in doing just this: it decontextualizes the relevant lexis of the article.

2.3.2. Making collocations more visual: Collocate analysis on AntConc, GraphColl and Voyant Tools

For collocations, the present study utilizes three tools to make them more visual. These tools are *AntConc*, *GraphColl*, and *Voyant Tools*. The first two are dedicated for the purpose of studying collocates: *AntConc* with the traditional method of finding collocates from a concordance table, and *GraphColl* for the study of collocational networks. *Voyant Tools*, on the other hand, is a multipurpose tool that includes both options. It also includes other means of analysing texts and corpora, which are covered elsewhere, and utilized in the materials. AntConc uses colour to highlight collocates in the table of concordances, GraphColl utilizes a network of collocations with colour, and Voyant Tools utilizes multiple options for collocations: a network like GraphColl, a "correspondence chart" similar to a concordance table, and a two-word phrase table.

2.4. Compiling a corpus

Before the computer revolution and widespread utilization of personal computers, compiling a corpus was a strenuous, time-consuming task. Due to this, a corpus was typically compiled only when it was necessary to preserve and categorize texts that were perceived to have a particular value to the generations that follow. Therefore, in the late Middle Ages, from where the first corpora date, corpora consisted of religious, theological texts that were seen as having an indispensable value. The first known corpus was compiled by a monks: 500 monks guided by Hugh of St. Cher (Barnbrook, Zyngier, and Vander 2011: 1). Later on, corpora were compiled that consisted of the writings of prolific, well-known authors whose writings were seen as being invaluable cultural capital – often after the author him- or herself had passed away. For example, a complete collection of Shakespeare's

writings falls into this category. The Shakespeare corpus well demonstrates the benefit of compiling a corpus to the further generations: these texts are now a central part of the Western Literary canon. In *Voyant Tools*, it happens to be one of the two default corpora included – showing how the work of collecting and digitalizing those texts still has relevance.

Nowadays, compiling a corpus is simpler, but in the case of massive, extensive generalized corpora, by no means an easy task. But, as a flipside, the corpora that exist nowadays can include significant amounts of text and other data: more than can be read during the lifetime of one reader. Corpus tools exist for the utilization of these massive corpora, like *Brown's* or *Longman's*. Some of them can be utilized by the computer tools used in the appendix of the present study. In the context of EAP, *specific purpose corpora* are even more useful. Specific purpose corpora dealing with EAP include Michigan University corpus of upper-level student papers and SciElf. SciElf is a corpus of academic English texts by writers with ten different L1s compiled by researchers collaborating across borders (Rowley-Jolivet 2017: 6-7).

In English for Academic Purposes: the texts in the said corpus are similar to the ones that the target group of this study encounter in reading, and have to produce themselves. Moreover, rather than using readymade corpora, a personalized corpus is even better, which is the focus of the present study. Previous research has shown that personal, specialized corpora are beneficial in academic learning – at least at a certain stage in the students' studies (Lee and Swales 2006; Charles 2014). Even better, in the study by Maggie Charles (2014) a majority of students persisted on using their personal corpora, in effect becoming corpus linguists in addition to their own academic field of study outside linguistics.

2.4.1. Compiling a personal corpus

In *English for Specific Purposes*, ESP, a key challenge for teachers is to know which are the frequently occurring terms and word structures that occur in the field they are teaching: what is the most important vocabulary that their students will encounter later on when studying texts in their subject area (Römer 2011: 209). In many academic disciplines such corpora exist, but as ESP and EAP contexts can be extremely varied, it is often the case that a ready-made corpus or teaching materials are unavailable. In those cases, a teacher can compile a corpus themselves. An even more efficient approach than teachers making a specialized corpus is that students make one themselves: an activity that furthers autonomy and allows for efficient inductive learning. This exercise may have

limited use in secondary education contexts, but has proved to be extremely successful in the context of tertiary education.

For example, in an innovative study on an EAP corpus pedagogical course by Lee and Swales (2006), non-native doctoral students from various disciplines writing their theses in English compiled corpora for themselves and used a concordance program, WordSmith, to discover linguistic phenomena from it. Lee and Swales note that the fact that the students were already doctoral students – the highest possible academic level that can still be said to be a student – means that they already possess appropriate macro-level knowledge of their field and the texts in it, they have already been acculturated to them (Lee and Swales 2006: 57). The researchers suggest that students need more support on the macro level: on the level of genre characteristics, target audience and so on (ibid.). In another study, which measured the long-term use of personal corpora to support academic writing, EAP students with diverse disciplinary backgrounds had compiled personal corpora for themselves on a course dedicated to supporting academic writing with a personal corpus (Charles 2014).

Most, 70%, of the corpus users continued using ~~suing~~ their personal corpora (Charles 2014: 33). Also, most did not add any new material to it, e.g. they had not made any changes to the corpus one year after the course when they originally compiled it (Charles 2014: 32). Perhaps one reason for not adding new material is that the process of doing so is somewhat time-consuming with *AntConc*, where the data needs to be converted to a different format and preferably cleaned up, and the adding needs to be done by manually adding the new text to the bigger corpus text – file by file additions are impossible, unlike in *Voyant Tools*, where the process is far more modular and intuitive. Still, if one wants to create a corpus that is reliable and efficient with *Voyant Tools*, the cleaning up process is rather useful. Tellingly, in the study by Charles (2014), an *engineering* student used corpora far beyond the context of the course, and unexpectedly compiled a total of three different corpora for different contexts and disciplines and used them to compare the language use in them (Charles 2014: 34). She used these to contrast and compare the language

As is the case of much of the work done on Data-Driven Learning with corpora, both studies utilized concordance programs as the method of corpus pedagogy. There is nothing wrong with the using concordance programs as a sole method of accessing corpora, as they are efficient in uncovering the textual context and collocations. It needs to be noted that the experimental course by Lee and Swales (2006) was successful. Concordance analysis is a mainstay in corpus research and pedagogy with tremendous potential for lexicogrammatical analysis: of correct adverbs, of collocations and so on.

However, on a structural level, on the level of discourse, on the level of fitting the text to the context and genre, the applications are limited (Lee and Swales 2006: 57).

The study by Lee and Swales was based on a voluntary EAP course arranged for PhD students. It had a high dropout rate for the students: only 50% of the students who had enrolled continued past the first lesson, something that the researchers themselves found surprising (Lee and Swales 2006: 59). This highlights the need to make direct corpus utilization more appealing and convenient – the rate of dropping out, considering the level of the students and their initial voluntary desire to attend the course is indeed high. In addition, previous research on concordance use cautions against "concordance burnout", an overuse of concordance programs as the sole method of deriving linguistic data from texts (ibid.). The concern for "concordance burnout" is valid: even corpus researchers themselves note that concordance analysis is extremely time-consuming (Baker 2010: 21). To avoid the burnout, the present study seeks to utilize less traditional methods of corpus-based data driven learning in addition to concordance tables: tools to visualize text and the collocational network in it.

The students in the course described in the study by Charles (2014), in addition to concordancers, used *Plot tools* so that words could be analysed by clustering them, providing one way to escape the concordance burnout. Since it was successful in fostering long-term use of corpora tools, there is empirical support for the idea to use multiple tools and methods of corpus analysis for a long-term benefit. In addition, the most often mentioned reasons for non-use: insufficient ease of use, inconvenience, and perceived slowness (Charles 2014: 35), why 30% did not continue the use of a corpus suggest that an easier, more convenient, and faster tool such as Voyant Tools, might support an even higher retention rate in the use of corpora. Particularly since one inconvenient factor that the students noted was that that the programs required to use the corpora had to be installed anew when working on a different computer (Charles 2014: 35). This is not the case with Voyant Tools, which is online – but a student using it still needs to access the corpus file created with Voyant Tools to be able to use the corpus. Fortunately, there are good cloud services available that can be accessed.

Being able to create a corpus for oneself confers many benefits both from the viewpoint of language learning, finding out content information from the text, and research. Of course, *any* collection of texts can be seen as a corpus – thus, a collection of articles a university student has gathered for the purposes of writing a thesis or an essay as part of a course can be seen as a corpus. However, these 'corpora' are not necessarily accessible to corpus tools – they have to be converted into a different format first.

Fortunately, making a corpus of texts of one's own choosing is a surprisingly simple task: step-by-step instructions for doing this with Voyant Tools, saving this corpus and utilizing it further for linguistic and discourse-level analysis are included in the appendix of the present study. These may be utilized by both teachers and learners. Some of the benefits that making a corpora of the texts one is studying, include being able search for relevant information from all of them instead of just one with search functions, the ability to compare the texts, and searching for concordances of relevant terms one is studying them, among others. These possibilities are covered in chapter 4 in more detail and utilized in the appendix.

2.4.2. Compiling a corpus of learner texts

In corpus pedagogy, teachers may want to create a corpus of learner texts to find out frequently occurring mistakes in them, so they can be addressed in further teaching. In the present study, this is mainly addressed in learning tasks, where students compile a mini-corpus of their own texts so they can notice the mistakes they do, overly repetitive phrases or terms in their own writing, and developing their *style* to fit the register and the genre they are writing in better. Therefore, the options teachers have in compiling corpora of learner texts do not require much attention in the present study. Previous research has shown that targeted intervention based on learner corpora by teachers can be effective (Römer 2011). Learner corpora can also be used to study interlanguage, and utilized as a part of Action Research pedagogy, a teacher-led reflective pedagogical approach (Tono 2009: 184-185).

The present study does not focus on corpora of learner texts, but some of the tasks in the material ask the users to contrast and compare their own texts with those of professional writers. There is even an option for the users to compile a corpus of their own texts and use them to reflect on their development, and then write with more awareness of their own style. By doing this, the learning is more constructivist and reflective: there is the back and forth between thought and action that is essential in constructivism (Tourmen 2016: 10). In the next chapter, I look at what a corpus of academic texts reveals: what are the distinctive characteristics of academic language and academic English.

2.5. Academic English as a subset of English

Academic English, or English for Academic Purposes (EAP), is utilized globally by English-native speakers, and ESL/EFL student users in English-medium instruction, and by professional academic writers in all disciplines writing to wider, global audiences (Mauranen 2010; Rowley-Jolivet 2017). Indeed, academia as a whole is "thoroughly dependent on cooperation across national boundaries" (Mauranen 2010: 7). It should be noted here that the term EAP is not used by all researchers, with some opting to simply refer it as a subset of *English Lingua Franca*, Academic ELF (Mauranen 2010: 6). Perhaps, since the field is so varied, it is more appropriate to speak of *Academic Englishes*. This 'pluralisation' of the term is similar to the corpus driven approach to English as a Lingua Franca: *Global Englishes* (Cogo and Dewey 2012: 21-22). The term *Academic Englishes* is also inspired by James Gee, who cautions against seeing school language and academic language as distinct, decontextualized varieties of English (Gee 2001: 63; quoted in Faltis 2013: 11-12).

For the purposes of this study, an important distinction should be made between "school language", which suggests the language used in primary and secondary education, and "academic language" of tertiary education in Universities and polytechnics. It should also be noted that the "styles of language" Gee refers to above corresponds more with *genre* in subsequent analysis, not *style* as the term is used in the present study. However, the crux of the argument remains: it is not possible to view academic English as a monolithic entity. Rather, it consists of many genres, subgenres, and contexts of use. Hence, while the main focus of the material package is fostering communicative competence in English for Academic Purposes in a way that is transferrable to and within the broad area of Academic English, or fostering the skills of aspiring writers, I view those skills as being genre-specific: they might not be transferrable to writing contexts outside EAP.

The viewpoint of writing as a communicative skill is labelled the *Skills Viewpoint* (Ivanic 2004: 222,225, 227-229) Here, it is supplemented by the *Genre Viewpoint* (Ivanic 2004: 225,232-234) on writing, and to an extent, by the social practices viewpoint (Ivanic 2004: 225, 234-237). These viewpoints, skills, genre, and social practices, cover writing from the viewpoint of the written text itself in the case of skills, and from the viewpoint of the writing event in the case of the genre viewpoint and the social practices viewpoint (Ivanic 2004: 225). The process discourse on writing (ibid.), is something that the inclusion of SFL seeks to address. Skills can also be labelled *communicative competences*, and in this labelling, the distinction between skills and social practices

becomes less clear: skills within the context of certain social practices, and communities of practice (Pyrko and Dörfler 2017) such as academia.

Communicative competence in L2, including EAP, consists of two distinct competences: *interpersonal competence* and *academic competence* (Saville-Troike and Barto 2017: 105-106).This division, and the role of academic language in it reflects a *competence-based* viewpoint: academic language is seen as a skill to be acquired (Ivanic 2004: 225, 227-229). Looking at the topic more from a more linguistic standpoint, the two types of language can be termed *social* and *academic* language (Faltis 2013: 12-13). These competences, or having skills in these two types of language, are not strictly separated from each other, as socially-oriented language is needed for full participation in classroom activities (Faltis 2013: 11), and, in a broader sense, in the academic community. Nevertheless, this study focuses on academic competence, despite the obvious importance of the interpersonal, insofar as these two can, in fact, be separated from each other as the writing, reviewing, responding, and peer reviewing of articles can be seen as taking part in the academic community of practice.

A caveat is in order: this study focuses solely on written academic English. Biber (2006) summarizes the research on spoken and written academic *registers*, or types of words and word constructions typical in contexts. The physical mode of production, that is, spoken or written, is the most significant source of variation (Biber 2006: 214). It should be remembered that the following, and the present study in general, applies only to written academic English, or EAP in writing. Further, it deals mainly with the types of texts students are expected to produce, e.g. assignments and dissertations, by contrasting them with research articles, text books, or more proficient dissertations. This leaves out many types of written academic language, many *Academic Englishes,* such as course management related writing, advising and institutional writing, and grant writing, which are all commonly occurring types of academic writing with their own distinct characteristics (Biber 2006: 215-217). It should also be remembered that not all research articles in English are necessary stellar examples of well-written English, nor are all texts produced by students limited to the novice level of proficiency in English. Keeping that in mind, the contrast here is nevertheless between a novice user and an advanced user.

From the perspective of L2 learning, these competences, academic and interpersonal, share many similar features: language is initially learned in chunks (Schmitt 2010: 139-140; Saville-Troike and Barto 2017: 148); linguistic, psychological, and microsocial contexts matter (Saville-Troike and

Barto 2017: 109); there is a division between spoken and written – although academic competence is generally seen as much more focused on written communication and interpersonal on spoken. As mentioned earlier numerous times, the focus of this study is on written academic English. In the following sub-chapter in this part of the thesis, I explore the core characteristics of academic written English: the *register* of EAP, the *genre* of EAP, and the *style* of EAP.

2.5.1. Register, genre, and style in EAP

Register refers to the language variety of the text type in question (Biber and Conrad 2009: 6). For example, in this study, the language variety is academic English, and the words and lexical bundles that occur in it can be said to belong to the *academic register*. Genre refers to a "Conventionalized categories and types of discourse" (Saville-Troike and Barto 2017: 211), in the case of EAP, genres that are classified under it include research articles, books that teach an academic subject, student essays and theses, to mention some written genres only. The last category, style, refers to the characteristic linguistic features associated with a sample of texts from a variety of texts, typically within the same genre or from the same author (Biber and Conrad 2009: 10, 18). Register and style are similar, the difference being that register features are functional, while style features are aesthetic (Biber and Conrad 2009: 54). The organization of textual features into register, genre, and style (Biber and Conrad 2009) is not in a hierarchical order: the study of register and style both deal with lexicogrammar, whereas genre is a more global category, (Biber and Conrad 2009: 50), or a more macro-level category when thinking taxonomically (Halliday and Mathiessen 2014).

For a novice user of EAP, the personal style is often still something that is developing, and rarely corresponds with typical samples from EAP genres of professional users of EAP, e.g. researchers. Of course, it is neither necessary, nor desirable, for aspiring L2 EAP users to lose their own distinctive style and uncritically adopt the style of advanced L2 users in entirety. Instead, the practical focus here is on describing and distilling the key distinctive features of the EAP register, so that novice writers can harmonize their own style to the academic genre of their choosing. In other words, the goal for aspiring writers is to fit their style of writing to the genre. To illustrate the relationship genre, register and style have in EAP and what is the necessary area of improvement for a novice user, I have created a table based on with the terminology from Biber and Conrad (2009) with supporting, empirical evidence of the differences from Scott and Tribble (2006: 132-158).

Table 2: Genre, Register, and style in EAP, novice and advanced writer. Based on Biber and Conrad (2009), and Scott and Tribble (2006: 132-158).

Aspect ↓	**Level**	**Novice user**	**Advanced user**
Genre:	Global. Level of the entire discourse	*Genre markers* present, occasionally inappropriate	*Genre markers* present and appropriate
Register:	Local. Lexico-grammatical and functional, focus on lexis	Words and lexical bundles, or *register features* sometimes inappropriate to the genre, *register markers* might be lacking	Words and lexical bundles, or *register features* varied and appropriate to the genre, *register markers* used appropriately
Style:	Local. Lexico-grammatical and aesthetic, lexis and grammar	May be highly aesthetic, *style features* personal, but sometimes off-genre or repetitive	May be highly aesthetic, *style features* personal, but fit the genre, or even define and reshape it

There are three italicized terms in the table that need explanation: 1) *Genre marker*, 2) *Register marker*, and 3) *style feature*. Biber and Conrad (2009: 54) define genre markers as "distinctive expressions and devices that are used to structure a text from a particular genre". They typically occur in pre-determined stages of a text, not throughout it like register markers do (Biber and Conrad 2009: 69-70); a typical marker in EAP might be the work "abstract" in the beginning of an academic article. The example genre marker immediately lets a reader familiar with the genre know that the text is likely a dissertation or a scientific article, and the part that follows the marker presents a condensed overview of the whole article, the abstract. Register markers are typical words or lexical bundles that occur throughout the text in a particular genre, they are pervasive and frequent (Biber and Conrad 2009: 53). Register features are similar, but also occur in other types of texts. However, in the particular genre where they are considered characteristic, they are pervasive and frequent, whereas register markers are exclusive to the genre they mark (Biber and Conrad 2009: 55).

In EAP, register features consist of single words typical in academic texts, such as the ones in the Academic Word List (Coxhead 2000), and lexical bundles typical in academic writing (Biber et al. 2004; Scott and Tribble 2006: 131-158). In addition to words and lexical bundles the distribution of word types like pronouns and nouns is also a register feature (Biber and Conrad 2009: 56). For

example, pronouns are rather rare in academic texts, and even the use of "I" to denote something that the researcher has done, or is doing, has traditionally been eschewed in academic writing, as it is seen as something subjective and based on opinion, rather than facts. However, the critical, postmodern trend of emphasizing agency, and thus acknowledging the subjectivity of the researcher, and opposing "self-appointed experts" (Pyrko et al. 2017: 391) has led to increased use of "I" construction in academic writing, particularly in the Humanities. Register markers, on the other hand, typically consist of terminology denoting concepts that are relevant to the academic discipline. Finally, style markers resemble register markers in the sense that they occur throughout the text and also include lexical bundles. Although genre limits the amount of stylistic choices available to the writer, but style markers do not serve the function of identifying the genre is question. Style markers, therefore, are associated with the *idiolect* of the writer: the distinctive, personal style of speaking and writing everyone has.

There are many facets of academic knowledge and academic writing where novice L2 users differ from expert academicians who use English as L1 or as L2 on an advanced level. It is not reasonable to expect that any simple solution would address all of those – only sustained practice and effort may do that. However, the misuse of terminology and context-inappropriate, or overly repetitive, use of lexical bundles are a part of EAP where a targeted intervention has potential to succeed, particularly at the level of register. West's General Service List from 1953, covering 2,000 most widely useful word families in English (Hyland and Tse 2007: 253) and the Academic Word List (Coxhead 2000) might be of limited use, as the use of words differs between academic disciplines greatly. In the present study, this is one of the reasons why these lists are not utilized for the teaching of EAP. In his seminal work on University Language, Douglas Biber (2006) studied the patterns of word use across disciplines both in spoken and written English University Language. The focus here is on written University Language, or EAP, and mainly from the perspective on what is useful for L2 users of EAP whose native language is Finnish. However, many of the characteristics of EAP described below apply regardless of the L1 in question.

2.3.2. General linguistic features of Academic English

For a student entering tertiary education fresh from High School, or otherwise with no background in the Academia, adjusting to University Language presents many obstacles to overcome. There are lectures, course books and articles assigned for reading that the student needs to understand (Biber 2006: ix), and essays and assignments that the student has to produce. Of course, there are non-

linguistic challenges here that could pose far greater challenges than the linguistic ones, especially in Natural Sciences: the student might not have the prerequisite talent for the field. In many fields, however, the linguistic challenges are closely intertwined with the academic, cognitive ones (Nagy and Townsend 2012; see also chapter 2.1.6). The linguistic barrier exists regardless of the language of instruction, as academic vocabulary and discourse are rather specialized. It might even be that that the 'linguistic shock' of academic language is greater for L1 students due to the fact that explicit attention to academic language in L1 English instruction is often insufficient (Biber 2006: ix).

In contrast, in ESL/EFL teaching in secondary and especially tertiary education, it is an area of explicit focus. This is due to the practical need to cope with EAP in university studies. There is explicit language teaching with an academic focus for students, so that the 'linguistic shock' of adjusting to EAP is lessened: these are typically part of the curriculum in the first years of study. They serve the purpose to acculturate students to the language during lectures, in course books, and articles assigned to them during their studies (Biber 2006: ix). In addition to the vocabulary and its link to the thinking required at that level (Nagy and Townsend 2012: 91), complex academic *discourse* is the one that poses a challenge (Biber 2006: ix).

It bears remembering that the general prior experience students have varies significantly: for example, the immediate impression that many readers have when reading a textbook on scientific subject, which is that the sentences or words in a textbook are long, is based on prior experiences with other types of texts (Biber and Conrad 2009: 52). Students that are novices in their field might already be experts in a different field and are doing their second degree. Or, in the case of readers who have encountered texts that have similar features the impression might be different (ibid.). For example, an L2 reader who is familiar with a longer sentence construction used in of English-language broadsheet newspapers will likely not gain an impression that there is anything particularly 'long' in the words or sentences in academic textbooks. Such a reader can likely follow the thinking presented in them due to the similarity in style. Academic writing and newspaper texts share some characteristics, but for comprehension, newspaper articles can be either skimmed or read carefully, while whether academic texts can be skimmed depends on the reader: some readers can either skim the articles or read the articles for comprehension, while other readers have the option of only reading them carefully (Biber and Conrad 2009: 111). Presumably, this difference is caused by familiarity with the topic area in addition to differences in aptitude. Therefore, while the newspaper reader has an advantage in knowing the style, when reading academic articles they are still dealing with texts belonging to a different genre with a different register.

Finally, academic language shares the common characteristic of language from other contexts, such as polyphony and intertextuality (Mesthrie 2009: 320). What makes academic language distinct from other forms of discourse is the expectation that the *voices* in the polyphony, and the intertextuality in it, become visible to others by following a strictly delineated procedure of quotations and citations. Simply put, unlike most other forms of discourse, which are implicitly intertextual, academic written discourse is explicitly intertextual.

2.5.2. Discipline variations in Academic English

Academic English language varies between disciplines in the level of register, the vocabulary employed in them (Biber 2006: 226; Hyland and Tse 2007; Nagy and Townsend 2012); in the use of passive voice vs. the use of pronouns such as "I" to express agency (Biber 2006: 226); and in the types of lexical bundles employed (ibid. see also chapter 2.2.4. on the classification of lexical bundles). In addition, there are differences in the focus of the texts. That is, whether they are procedural, content-focused, narrative or non-narrative (ibid.). Or, to view it from a slightly different angle: there are discipline variations in the types of socio-semiotic processes that academic texts in different disciplines serve (Halliday and Mathiessen 2014: 37; see also Figure 3 in the next chapter for the classification of the socio-semiotic processes). Despite the fact that disciplines may have internal variation in their language, research on discipline variations commonly divides the disciplines into three (Hyland and Tse 2007: 240) to five (Biber 2006: 225-227) different disciplines for the purposes of studying discipline variations.

In the case of Hyland and Tse (2007: 240) the three disciplines in which variations were sought were Engineering, Natural Sciences, and Social Sciences (Hyland and Tse 2007: 238-239). In their division, *applied linguistics* was counted under Social Sciences, as was *business studies* (ibid.). In contrast, Biber (2006: 224-227) had separate categories for Humanities and Business. It was found out that Humanities and Social Sciences share many similar features (Biber 2006: 224). Therefore, the inclusion of applied linguistics, which could be categorized under humanities in the study by Hyland and Tse (2007) means that the findings of the study utilizing three categories conform with the one with five categories on that part. However, business, as a broader category not classified under social sciences was found to have more in common with engineering: forming a pole opposite to the Humanities and Social Sciences, with Natural Sciences somewhere in the middle (Biber 2006: 224). This means that due to the inclusion of business under social sciences in their three-category

classification system (Hyland and Tse 2007: 238-239), the studies do not conform with each other when it comes to Business.

However, the study by Hyland and Tse (2007) analysed vocabulary on the single word level, the study by Hyland (2012) on the level of lexical bundles. The classification by Biber (2006: 224-227), as mentioned, covered lexical bundles, and the dimensions of procedural/content-focused and narrative/non-narrative on a textual in addition to the type of vocabulary employed. To demonstrate the differences, I have adapted the findings and tables of the both studies into a table format: because they do not conform completely with each other in the classifications of disciplines and study different dimensions of language, the findings are presented separately but combined in one table. Note that there is no 'rare' option in the features: academic language is diverse and varied in all disciplines, some features are simply more common in others.

Table 3: Discipline variations in register, lexical bundles, and focus in written Academic English (adapted from Biber 2006: 225; Hyland and Tse 2007: 240; Hyland 2012: 163).

Study→	**Biber 2006**						**Hyland and Tse 2007; Hyland 2012**			**2007 /2012**
Discipline→	Bus.	Eng.	NatSci.	SocSci.	Hum.	Disc. →	Eng.	NatSci.	SocSci. Hum.	2007 2012**
Dimension of analysis↓						Dim. of Analysis ↓				
Diversified vocabulary	0	0	+	++	++	Diversified vocabulary	0*	+*	+*	2007
Specialized vocabulary	0	0	+	++	++	Specialized vocabulary	0*	+*	+*	2007
Abstract/process nouns	++	++	0	+	0	*Process, Design, function*	++	+	+	2007
Concrete/technical nouns	0	+	+	0	0	*Research, Strategy*	+	+	++	2007
Referential lexical bundles	+	+	++	++	+	Referential	++	++	+	2012
Intangible framing bundles	+	0	0	+	++	Intangible framing	0	0	++	2012
Place referential bundles	0	0	++	+	0	Place referential	0	+	0	2012
Epistemic stance bundles	+	0	0	+	0	Epistemic stance	0	0	0	2012
Ability stance bundles	0	+	+	0	+	Ability stance	+	+	0	2012
Importance stance bundles	0	0	0	+	0	Importance stance	0	0	0	2012

Text: Procedural	+	+	0	0	0	-	-	-	-	-
Text: Content-focused	0	0	++	+	+	-	-	-	-	-
Text: Narrative	0	0	0	+	+	-	-	-	-	-
Text: Non-narrative	0	+	+	0	0	-	-	-	-	-
Passive voice	0	0	+	+	0	-	-	-	-	-

Explanation:
Bus: Business Eng: Engineering; NatSci: Natural Sciences
SocSci: Social Sciences Hum: Humanities
0: Low to normal frequency
+: Very frequent
++: Extremely frequent
-: Not studied
Italics refers to words: the study by Hyland and Tse (2007) measured the frequencies of individual words that are common in general academic vocabulary in different disciplines.
*: Based on the coverage of General Service List and Academic Vocabulary List: low coverage of common, shared academic vocabulary suggests more diverse and specialized vocabulary
**: Based on analysing select words from the table of most common lexical bundles per discipline (Hyland 2012: 163)

As seen in the table above, the fields of Business and Engineering are rather similar to each other when it comes to the type of vocabulary employed in them – although the content of the vocabulary, the themes covered in it, are different. Namely, they contain a high amount of *Abstract/Process nouns* (Biber 2006: 225). Examples of these nouns are process, design, and function (Hyland and Tse 2007: 240). A similar focus is true on a textual level: business and engineering texts are procedural. On the other extreme, Humanities and Social Sciences contain a particularly high amount of *Specialized Vocabulary* as well as *Abstract Vocabulary* (Biber 2006: 225). On a textual level, they are more narrative than other types of texts (ibid.). When it comes to lexical bundles, the most notable difference is the preponderance of *intangible framing bundles* in the Humanities and *place referential bundles* in Natural Sciences (Biber 2006: 225; Hyland and Tse 2012: 163). This makes sense: Natural Sciences deal with the precise identification, categorization, and location of natural objects and phenomena, while Humanities deal with abstract subjects. The relation of these abstract subjects is expressed via intangible framing bundles.

The basic structure of the table is from Biber (2006: 225), with the study by Hyland and Tse (2007) on discipline variations in academic vocabulary, and a part of the study by Hyland (2012) on the variations on 4-word lexical bundles between disciplines (Hyland 2012: 163). It should be noted that the table is not a perfectly accurate representation of the latter two studies: in the case of Hyland and Tse (2007), the diversified and specialized vocabulary is an estimate. Also, in the case of nouns, two representative ones were chosen. The same was done with the lexical bundles from Hyland (2012): a rough summary is presented here based on the frequency of certain bundles that belong to the

categories identified by Biber (2006). For example, '*in relation to the*' is an intangible framing bundle. Macro-level features, e.g. whether the text is content-focused or narrative are solely from Biber (2006).

In summary, discipline variation in EAP is a significant factor. Often, it is not the case that some features are entirely missing from disciplines that are common in others, but that the frequencies vary due to different norms and purposes that the texts in different disciplines have. Particularly noteworthy is the diversity of vocabulary in Humanities and Social Sciences, which suggests that explicit focus on academic vocabulary is especially important in these disciplines.

2.5.3. The processes of Academic English

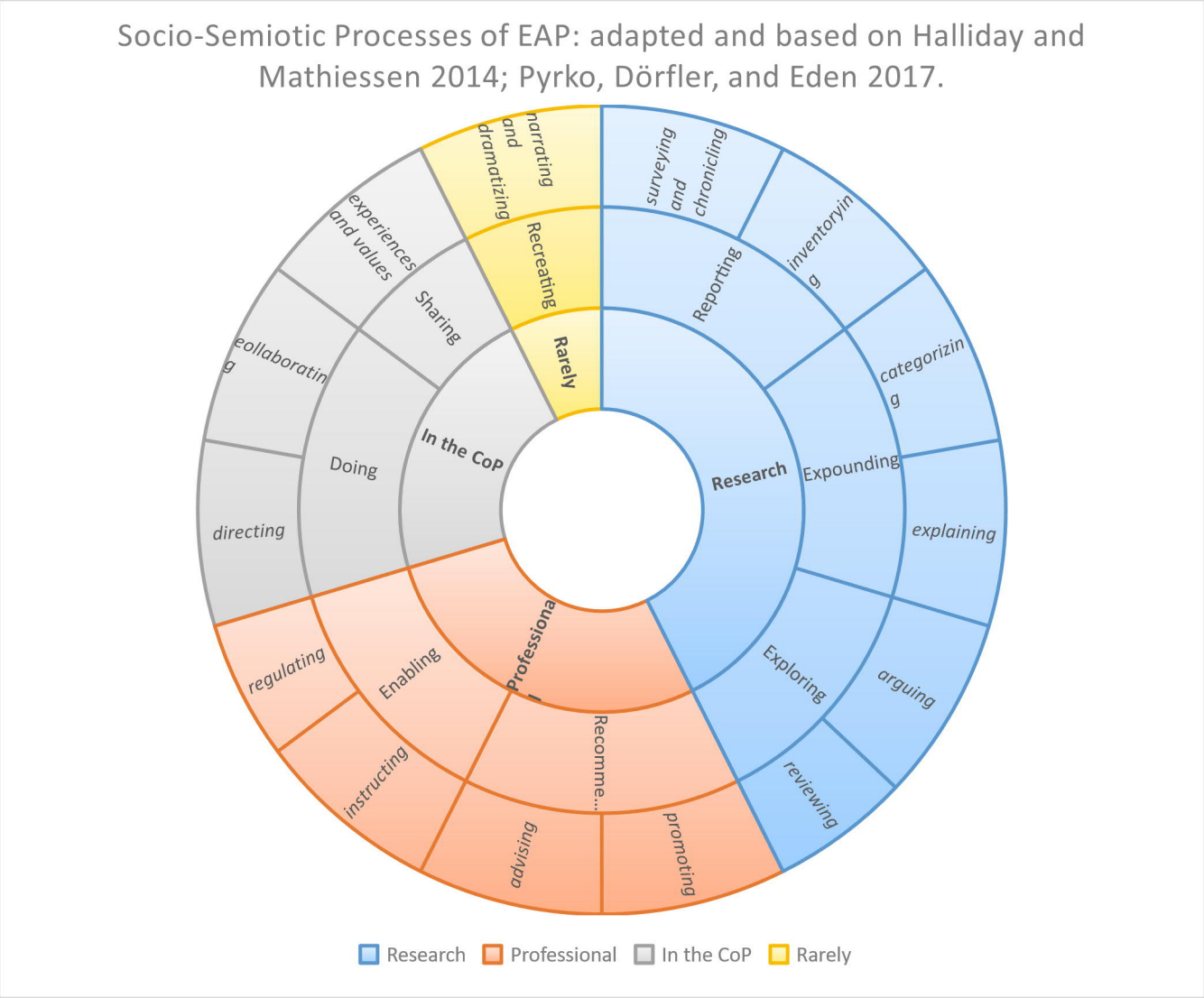
Even though the present study focusses on the Academic competence aspect of language use rather than the interpersonal one (Saville-Troike and Barto 2017: 143-144), the *communities of practice* viewpoint, which is utilized here, forces a partial reconsideration to the division. The division, broadly speaking, is a valid one – but to maintain consistency with the communities of practice viewpoint with the spectacular categorization options conferred by Systemic Functional Linguistics, it is beneficial to look at how academic communication, even in its most traditional, strictly *non-interpersonally* perceived domains such as research articles and theses fits into *socio-semiotic activity types* (Halliday and Mathiessen 2014: 35-38) of communicative text types: what sort of an activity they represent. That is, whether it is to inform, to promote and so on. In full, there are a total of eight functions for a text, which are later divided into sub-categories (ibid.):

1. Reporting,
2. Expounding,
3. Exploring,
4. Recommending,
5. Enabling,
6. Doing,
7. Sharing, and
8. Recreating.

Arguably, even the most traditional academic texts serve as activities beside pure research: they are part of the profession of a researcher, and they represent engagement with other researchers in the

community of practice of academia. Therefore, I have synthetized these three purposes with the eight theme functions of SFL in the figure below:

Figure 4: Socio-Semiotic Processes of EAP. Adapted from Halliday and Mathiessen 2014; additional category (the innermost ring) based on Pyrko, Dörfler, and Eden (2017).



The figure above, when it comes to the two outermost circles, is almost an exact replica of Halliday's and Mathiessen's (2014: 37) figure that categorizes the *socio-semiotic*, that is, the social purpose and the meaning assigned to various forms of written communication in categories. The categories are not mutually exclusive: a text may be *Enabling and instructing*, *Expounding*, *explaining and categorizing*, and *Doing and collaborating* at the same time. The middle circle represents the broader categories, such as *Enabling*, *Expounding*, and *Doing*, while the outermost circle represents sub-categories within the broader categories: an *Enabling* text can be *regulating* and/or *instructing*.

The innermost circle is an addition, which utilizes the notion of *community of practice* (Pyrko et al. 2017), and makes the categories even broader. Hence, *Doing* and *Sharing* are categorized as being 'sub-sets' of the category *In the CoP* – CoP here refers to the community of practice. *Research* texts generally fit in the categories coloured blue, while texts that are part of the *profession* are coloured orange: they represent texts such as course books and course materials, in addition to other pedagogic written communication. The *'In the CoP'* category refers to an implicit purpose that most text, even research texts serve: they are ways of *Sharing* and *Doing* things with the community of practice. Finally, the *Rarely* category exists to classify the text purposes that academic texts less often serve: *Recreating* events, or *Dramatizing*. However, when studying texts across disciplines (Biber 2006: 226), it has been found that in the Humanities, texts are often narrative (See table 3). Thus, they can be categorized as being recreating. This is certainly true in fields such as ethnography: where the researcher seeks to recreate their experiences of the field-work as faithfully to the original experience as possible. Note that the categories of what is related to the profession and the community of practice can be, in practice, rather overlapping: of course the professional functions occur in the community of practice, and of course part of the profession is engagement with the community of practice.

3. Knowledge Building and Pedagogy

In the previous sections of the thesis, I explored corpus linguistics as a field of study in the field of linguistics, and in it, I presented an argument on why it is relevant in higher education. It was necessary to distinguish corpus linguistics from other approaches and explore the view of language it presents. In the pedagogical part of this thesis, I search for a model of knowledge, a model of knowledge that is congruent with corpus linguistics, explore the difference between theoretical and practical knowledge, personal and shared knowledge, and take into account other factors affecting language learning. These factors include behavioural, cognitive, and psychological factors. Together with corpus pedagogy, they form the basis for the choices in the materials presented in this portfolio thesis.

3.1 Dimensions of knowledge

Science and creation of academic knowledge is about preserving, transmitting, and uncovering knowledge. Therefore, it is necessary to understand the *nature of knowledge*: what it is, how it is transmitted, how it is created, and what types of knowledge there are. In other words, it is necessary to explicate the epistemological basis of this study: how it informs the decisions I have made in

producing the materials. The materials have a dual purpose of language development, and the development of the users as purveyors and producers of academic knowledge: it is not possible to separate the use of academic language from academic knowledge (Nagy and Townsend 2012). Further on, I explore views of knowledge that emphasize its role in building communities: in this case, the community is the international academic one.

3.1.1 Epistemological Basis for Corpus Material Pedagogy

In Chapter 2.2.2. I explored the question on the levels of knowing a word and concluded that the focus of this material package is mostly in increasing the depth of word knowledge, at least when it comes to concepts. This is especially true when dealing with words and terms that convey an understanding of theories, such as *discourse*, or with abstract concepts, like various *–isms*, such as *nationalism* or *socialism*, which can be the focus of academic studies by themselves, in Humanities and Social Sciences. Understanding concepts like these is not so much about L2 learning, than it is about building content knowledge in the relevant field: building connections, making distinctions between related concepts, and understanding the possible controversies in the use of the terms by different users. Consequently, I review models of knowledge.

Three models of knowledge are relevant here: *empiricism, constructivism,* and the *probabilistic* one. First, empiricism is valid here due to its focus on observation and opposition to *a priori* knowledge and intuition. An empirical approach is what characterizes the approach of corpus linguistics to the study of language: as discussed in Chapter 2.2., corpus linguistics is descriptive, the 'rules' concerning the *lexicogrammatical* organization words in sentences are derived from corpora, rather than being something that prescriptively exist *a priori*. Similarly, intuitive guesses about the usage of words, or formulaic sentences, are often wrong, or if not wrong, at least less accurate than data derived from corpora. Naturally, this is not to discount altogether the use of native, or near-native in the case of advanced users, intuitions about words: utilizing corpora in all communicative contexts is implausible due to the constraints of time and the availability of tools.

Second, constructivism offers a plausible account on the development of knowledge for an individual and is necessary for more advanced stages of learning a language, such as understanding abstract concepts conveyed with language. For Lev Vygotsky, the role of concepts is central in organizing the thought processes of a person (Vygotsky 1931; quoted in Pass 2004:67). While modern psychological research may see the relationship between language and thought in a more nuanced way (Pinker

2007), the role of language in making the thoughts of a person known to others cannot be questioned. Also, constructivism is in line with the understanding of L2 learning and pedagogy presented earlier, with its focus on *action* (Tourmen 2016: 10), e.g. engagement with the observed knowledge and connection-building, which is central in vocabulary learning.

In addition to the individual account, the construction of knowledge socially, *social constructivism*, describes the notion of academic knowledge building in EAP as being situated in its community of practice (Pyrko et al. 2017). It also empowers the learner to develop their own understanding in their field by taking into account the relation of the semantic knowledge in the learners mind, and the corpora with the concept of *constructing* knowledge. Even if the tasks presented here are more individual, the learner still engages socially with the broader academic community, at least via reading. Of the two major constructivist thinkers with a major impact on pedagogy and developmental psychology: Lev Vygotsky and Jean Piaget, Vygotsky's ideas seem to be better applicable to the field of L2 pedagogy, or language learning in general. However, their viewpoints can be combined (Pass 2004: 107-117).

Third, there is the probabilistic model of knowledge, which is in line with new research in neuroscience (Tenenbaum, Kemp, Griffiths and Goodman 2011; Pouget, Beck, Ji Ma and Latham 2013). Consequently, it has had a major influence on cognitive linguistics and psycholinguistics. Empiricism is the viewpoint that is closest to corpus linguistics, and the linguistic data a learner is working with; constructivism gives a plausible model and a blueprint for the accumulation of knowledge towards higher order cognitive processes; and the probabilistic model is the one that explains why people make the choices they do, by analysing the cognitive processes that occur in decision-making. In language learning, the probabilistic model presents a plausible account of language learning, which is in line with the information processing framework of the present study.

Probabilistic language learning theory is closer to generative as opposed to discriminative (Hsu, Chater and Vitanyi 2011: 380-383). In generative theories, language is learned naturally from input, whereas discriminative models language is learned by making categories and distinctions of right and wrong utterances. Probabilistic language learning theories differ from natural language learning theories, where no categories are needed, by being concerned with the *probability* of occurrence of given utterances, which then shape the regularities of language (Hsu et al. 2011: 380), making it appear as there were categories. This approach is suitable with the statistical nature of corpus linguistics.

Of the epistemological viewpoints presented, the probabilistic one, and the Bayesian, or perhaps more accurately *the Laplacian* (Pouget et al. 2013: 1-2), after the mathematician Pierre Simon Laplace's concept of all knowledge being probabilistic, most accurately represents the data-driven, 'fuzzy logic' aspect of corpus linguistics (Boulton 2017: 1). It also aligns well with empiricism. In relation to constructivism, there are surprising similarities between the probabilistic viewpoint and particularly the constructivism of Piaget (Tourmen 2016: 11-12). Further, the probabilistic model offers a way to move past some of the classic dichotomies of cognitive science (Tenenbaum et al. 2011: 1285).

These dichotomies, "Empiricism vs. Nativism", "domain-general vs. domain-specific", "logic versus probability", and "symbols vs statistics" also appear in linguistics as schools of thought concerning the nature of language and language acquisition. First, the "Empiricism vs. Nativism" debate mirrors the debate of behaviourist such as J.F. Skinner against the arch-universalist Noam Chomsky, and in more modern times, the debate between modern proponents of universal language acquisition and corpus linguists. The second debate appears in linguistics in the form of how transferrable language skills are. The third question I have explored in relation to prescriptive vs. descriptive view of language, opting for a compromise: Systemic Functional Linguistics. Fourth, the symbols vs. statistics is similar to the approaches of semioticians and corpus linguists: symbols, and statistics respectively. For the fourth debate, the visualisation of corpus data serves as a practical bridge, and the probabilistic model a theoretical one.

The function of the probabilistic model in the present study is that it provides a model of knowledge that combines the insights of constructivism in the field of language development and learning with the empirical nature of corpus linguistics. Corpus linguistics, in turn, provides a bridge for the integration of the probabilistic model of knowledge into SLA pedagogy (Tono 2009: 189-191). In the case of the present study, the SLA framework is already something that can be integrated into corpus linguistics without difficulties: the information processing framework, in which learning is learning (Saville-Troike and Barto 2017: 77-78). The learning thus occurs based on processing the probabilities, not unlike in how collocations, for example, there is a probability for a word occurring with another, the strength of association. As a summary of the decision to combine these frameworks for the needs of the material package and the probabilities of a word occurring are corpus linguistics, while constructivism provides the theoretical basis of scaffolding the learning, providing a low-threshold entry to the use, and putting the theoretical knowledge into practice.

3.1.2. Types of knowledge: theoretical and practical

Modern societies are *knowledge economies* (Mokyr 2005). In this knowledge economy we currently live in, the ability to communicate internationally is of paramount importance. Currently, the language of international cooperation and communication is most often English. As a consequence, ELF in the business world is a topic that in all likelihood has been studied already widely, as back in 2010 it was a topic that was somewhat understudied but quickly gaining more interest along with studies on ELF in Academia (Mauranen 2010). ELF in Academia is, as mentioned earlier, referred to as English for Academic Purposes, EAP in this study. This chapter, however, is not about ELF *per se*. Instead, here I seek to analyse the accumulation of competence in from different perspectives: how it relates to knowledge economy by dividing into the two types of knowledge: the theoretical omega-knowledge of the *savants* and the practical lambda-knowledge of the *practicants*.

Although the notion originally did not occur in that context, the omega and lambda types of knowledge can be applied to corpus linguistics: with omega knowledge being the ability to understand the principles behind it and the lambda knowledge being the practical ability to use corpus tools. To complement these and to harmonize the notion of knowledge here with the constructivist notion of knowledge, I also utilized the *community of practice* model of knowledge (Pyrko et al. 2017) to analyse how the users of EAP practice thinking together (Pyrko et al. 2017: 390) in their development as users of that particular language variant. In linguistics, the distinction between theoretically-oriented and practically-oriented linguists has been strong particularly in the English-speaking world, with linguists being far more concerned in studying the fundamental questions of what language is and how does it operate in the mind.

The community of practice model of knowledge draws heavily from Michael Polanyi, for whom knowledge, in its richest form, can only exist as personal knowledge, the *indwelling* of knowledge (Polanyi 1968; cited in Pyrko et al. 2017: 392). Polanyi, however, was open to a more social view of knowledge, as it is possible for the indwelling of knowledge, that is, people's personal knowledge and expertise, whether theoretical or practical, to become "*interlocked on the same cue"* (Pyrko et al. 2017: 390). The meaning of the interlocking is that people share knowledge when working together on the same topic. Pyrko, Dörfler, and Eden (2017) see this in organizational communication. The same concept can be readily applied to knowledge building and sharing in academia: an academic discipline consists of people whose *indwelling* of knowledge is *interlocked* with each other. It is

through the reading and writing of academic texts that people both develop their indwelling of knowledge, and interlock with it with the knowledge others in the field have.

3.2 Cognitive and other factors involved in the learning process

The purpose of this section of the study is to explore how learning occurs, how both cognition and emotion are involved in it, and what sort of obstacles there are to learning. Further on, I explore how text visualisation, converting verbal data, at least partially, to visual, spatial, and numerical data can help as a support method to overcome obstacles to learning.

3.2.1 The effect of affect: the link between cognition and emotion in learning

The view that sociocultural theory has on cognition and emotion differs from traditional psychological approaches. In traditional approaches, cognition and emotion are seen as different domains, the domain of the mind and the domain of the body. In the Western culture, inasmuch as such a monolithic entity can be said to exist, the dictum 'mind over body' summarizes the attitude towards the bodily emotions. In contrast to this, the sociocultural theory sees the two as inseparable, and views *affect* as a key to the development of mind (Swain et al. 2010: 76). For Lev Vygotsky, emotions and cognitions are part of a dynamic system that is life, and he in his work, he sought to demonstrate the interconnectedness of emotions and cognition (Vygotsky 2000: 10; cited in Swain et al. 2010: 82). Current trends in psychological and neurological research appear to be vindicating Vygotsky's view at least partially, with particularly the stomach being considered 'a second brain' in its effect on moods and mental health, with the connection between gut and brain being bidirectional (Clapp, Aurora, Herrera, Bhatia, Wilen, Wakefield 2017). In language pedagogy, it has already been a widely accepted fact that emotional factors have a tremendous impact on the learning process, on the ability to understand a foreign language, and on the ability to produce it. This is especially the case in in primary education, secondary education, and special education – many of Vygotsky's insights stem from his experience as a teacher of Russian children with special needs in the 1920s (Swain et al. 2010: 16).

Personal stories from students attending higher education confirm the impact of emotions in learning, even in tertiary education. In a story relayed by Swain et al. (2010: 77), a graduate student of education named Grace was not able to understand the meaning of academic terms in her field of study, such as *literature review* or the specialized vocabulary in her field that was expected of her,

she did overcome this by attending extra classes (Swain et al. 2010: 77-81). The story illustrates that the effect of insufficient knowledge of academic English vocabulary not only impacts the 'success' of the student in her or his attempt to participate in the community of practice (Pyrko et al. 2017) and their thinking (Nagy and Townsend 2012: 101), but also their emotional state.

3.2.2 Cognitive factors

In chapter 2.1.5, in the part covering vocabulary in language learning, the topic of cognitive linguistics, and the contribution it offers to the material package was mentioned in relation to corpus linguistics. In this pedagogical theory part of this material package thesis, I explore the cognitive aspects further. First, there is the yet unanswered question on how, exactly, does language work in the brain (Pinker 2007). Second, there is the issue of cognitive variables between the learners and the users of this material package. On the first question, of particular relevance here is how words function in the brain, e.g. whether the human mind works via words and language, or whether words are linked to concepts, and how exactly this linkage happens. This is relevant for the tasks that cover the understanding of concepts. The second question, on the other hand, concerns the practical issue of taking into account the different types of learners who might be using the material package. Since the fundamental question posed by the first question remains yet fully completely answered by science, and would be far beyond the scope of an MA Thesis, let alone one that is only tangentially related to neurology and cognitive science, I utilize two existing viewpoint that are particularly suitable for the context of Academic English: *Radical Pragmatics* and *Conceptual Semantics.*

Steven Pinker (2007: 89-151), in a preamble to present his own view of language in the mind, Conceptual Semantics, presents an overview of the three other viewpoints: *Extreme Nativism,* Radical Pragmatics, and *Linguistic Determinism.* Furthermore, he claims that Conceptual Semantics is compatible with all three (Pinker 2007: 150). However, as already mentioned before in chapter 2 and in chapter 3.1, I reject in the present study, to a significant extent, *a priori* notions both on language, words, and knowledge. Therefore, in this study, only Radical Pragmatics and Conceptual Semantics are utilized of the four. However, to justify the choice, a brief summary on the other two viewpoints is in order. To recap, the viewpoints are, in chronological order of their discovery:

1. Extreme Nativism,
2. **Radical Pragmatics,**
3. Linguistic Determinism, and
4. **Conceptual Semantics**.

The viewpoints used in the present study are in bold font. Extreme Nativism is an approach that is in line with Universalist views of language, such as the one proposed by Chomsky (Pinker 2007: 30), and solidly on the side of 'nature' in the nature-nurture debate (Pinker 2007: 92). In extreme nativism, words have an innate meaning in the human mind: concepts are universal and exist as an 'innate inventory' *a priori*, as building blocks of cognition (Pinker 2007: 93). Furthermore, extreme nativism, to qualify as being extreme, makes a claim that most words cannot, in fact, be broken into their more elementary constitutients. For a pedagogical approach based on extreme nativism, the task of the language learner is then to find out the words that describe the concepts in their L1, subsequently in L2, and so on if more languages are learned. Some researcher see this viewpoint as one that is not an accurate description of how words occur in the mind (Pinker 2007: 98-107).

Here, the viewpoint is not utilized, because it is not in harmony with neither the pedagogy of the present study, nor does it seem particularly fitting with the view of language espoused by corpus linguistics, which is in direct opposition to nativist approaches. Therefore, it does not fit the context of this study. The approach of extreme nativism does not appear to be suited for L2, nor the learning of academic concepts. The precision of innate word senses required and in many cases demonstrated by extreme nativism, (Pinker 2007: 150) is not suited for the learning of polysemous words that are ubiquitous in academic texts. Also, the exact relation that words in L1 and L2 and the concept have with each other is still rather complex, as discussed in chapter 2.1.5. It is unclear how the innateness of concepts assumed by Fodor (Pinker 2007: 90-92) in Extreme Nativism would fit here.

Radical Pragmatics is based on the premise that words do not have an innate meaning, but they always derive their meaning from the contextual factors. In fact, for a proponent of Radical Pragmatism, innate meanings do not exist *at all* (Pinker 2007: 107-108). The idea fits particularly well with the concept of *polysemy*, which is central in EAP and corpus linguistics, the idea that words have multiple meanings. However, Radical Pragmatism differs from mere polysemy, even in the Bakhtinian sense where words always carry a taste of their previous contexts of use (Mesthrie: 176), as that idea presupposes that there are some cognitive, even innate meanings in play: it could be said that the innate meaning is modified by 'the tastes' a word acquires from their previous contexts of use. This process can be seen in the case of *pejoration,* the process where a previously neutral word acquires negative meanings due to it being used as an insult, for example.

The viewpoint of Radical Pragmatics seems to suit corpus linguistics perfectly. Radical Pragmatics is to the understanding of words in the mind as corpus linguistics is to the understanding of words in a text or a corpus of text: both presume a supremacy of the context; (Pinker 2007: 107-108) and the assumption that there is innate meaning in a word, only the one it derives from the context exist in corpus linguistics as a slightly weaker version. The main flaw of Radical Pragmatics for the purposes of this study is that it provides no recourse against the "semantic chaos" (Pinker 2007: 112) of total polysemy – the idea that words have no intrinsic meaning. In practice, it is also not true: there are words, particularly verbs, which language users use to the degree of precision that suggests an innate meaning (ibid.).

The third approach is that of Linguistic Determinism. Linguistic Determinism is an approach favoured in linguistic anthropology and closely linked to *Sapir-Whorf hypothesis*: the idea that the language a person knows and uses determines how they think. The strong version of the Sapir-Whorf hypothesis claims that the entire perception of reality, e.g. time and place is dictated by the language used to conceptualize them. In the strongest, purest version of Linguistic Determinism, the relation between thought and language is reversed completely: in it, language creates thought, not vice versa (Pinker 2007: 133-134).

A relevant aspect of linguistic determinism, for the purposes of the present study, is the vocabulary of colour – not all languages possess all the words for all colours, for example Korean does not have separate words for green and blue. However, languages always contain vocabulary to discriminate between light and dark, and the colour that all languages have a separate word is red: red is always the one to break away from the light-dark dichotomy. (Arnheim 2004: 331-332). Instead, for understanding reality, language is but a *medium*, a shared medium that is still insufficient in describing everything (Pinker 2007: 150). In addition, the claims made by proponents of Linguistic Determinism have a tendency to fall apart under closer empirical scrutiny, and they consist mostly of hoary myths about distant native people contrasted with Westerners who possess supposedly more advanced vocabulary, which then, again, is supposedly capable of describing reality more efficiently.

Unsurprisingly then, Linguistic Determinism is not utilized in the present study. The approach simply is not suitable for Academic English due to the previously mentioned *semantic chaos* it presupposes. In Academic English, and academic communication in general, this semantic chaos is the exact opposite what it aspires to be: as precise, nuanced, and as objective as possible. This is not to say that the postmodern critiques of 'expert hegemony' are without merit. They are not, as discussed in

chapter 3.1.3. Nor is it to say that critical studies on how we use language to refer to minorities and disadvantaged people are not important. Quite the contrary, those are goals and inquiries whose importance is not in question.

The fourth approach, and the second viewpoint used in the present study, is that of Conceptual Semantics. In Conceptual Semantics, the meaning of words and sentences exist as *abstract formulas* in the mind (Pinker 2007: 125). Here, this approach supplements the one offered by Radical Pragmatism particularly well: the purpose here is to further the understanding, the *indwelling* (Polanyi 1968: cited in Pyrko et al. 2017: 390) of the concepts. Therefore, the open-endedness of that Radical Pragmatism offers when it comes to understanding a word, or rather, the one that it does *not* offer, cannot adequately be seen as a pedagogic approach on its own. As an approach, it is even more unsuited for learners who are analytical and rule oriented (Ellis: 2008: 659-671) than incorrectly applied corpus pedagogy can be. To fill this gap, Conceptual Semantics serves a purpose. However, as the cognitive model used here is not completely in line of with that of Pinker – one main disagreement being that the *connectionist* model is something that he rejects, whereas for the purposes of this thesis, it is perfectly usable and particularly well in agreement with the textual level: with the idea of collocational networks. Nevertheless, the idea that words remain an incomplete representation of everything that the mind is capable of (Pinker 2007: 124) – the main theme of Conceptual Semantics – is something that I can fully agree with and utilize for materials teaching concept knowledge in EAP.

3.2.3. Overcoming psychological blocks to language learning

The stated purpose of the present study is the creation of materials that present a low-threshold entry to corpus linguistics for the purpose of learning EAP. In order to lower the threshold that students have in utilizing the tools presented here, to scaffold the learning process in a way that does not immediately alienate the learners, anxiety, the combined effect of pessimism, submissiveness, fear of failure (Bialystok 2002: 236), and a perception of low current ability must be taken into account. In constructivism, it has also been found out that fear narrows the *Zone of Proximal Development*, ZPD – the optimal level of challenge in language learning (Mahn 2008: 28; quoted in Swain, Kinnear, and Steinman 2010: 83). In the context of EAP, the learners likely already have some beneficial coping mechanisms to use, and have built resilience to cope with the feeling of anxiety. However, as addressed earlier, corpus linguistics is currently under-utilized by language learners. A possible

reason for this, in addition to the simple fact that there are other alternatives, is that due to the lack of exposure to corpus linguistics, the learning of a new method provokes anxiety.

Anxiety is a factor in language learning that most L2 learners find familiar. Anxiety can be divided into two types of anxiety: *state anxiety* and *trait anxiety*. State anxiety refers to the situation where a learner is frustrated by the demands of a particular situation: for example, a test that seems to be beyond their current abilities, or the perception of being made to 'perform' in the foreign language suddenly. Trait anxiety, on the other hand, refers to a relatively static psychological variable that varies between learners: some learners feel the effects of anxiety more and become anxious easier. (Bialystok 2002: 236). Psychological research distinguishes between two basic coping mechanisms, or self-regulatory strategies, that people have to deal with anxiety: strategic optimism and defensive pessimism. These strategies are strongly related to personality traits, so it is practically impossible for a defensive pessimist to become a strategic optimist. (Norem 2008: 124-126). Moreover, the performance of both types suffers when they adopt the preferred strategy of the other type (Norem 2008: 126).

Strategic optimism is an approach where the goal is first visualized, or otherwise elaborated on, and a strategy is generated beforehand for the success. The strategical optimist then creates a plan for success, and follows up on it. (Norem and Illingworth 2003: 352). For the strategic optimist, the expectation of success serves as a motivator and they typically expect high results (Norem 2008: 124). Failing to meet the goals fully generally does not bother a strategic optimist much, provided they have done their own part. As attractive as strategic optimism and positivity sounds, these terms are sometimes loaded, and optimism is highly valued, particularly in American culture (Norem 2009: 125), and the entire topic has a tendency to suffer from a 'conceptual mess' (Held 2018). They do not adequately address the benefits defensive pessimism and some negative-labelled approaches have, for instance in test-taking, and even more importantly that these traits can be innate (Norem 2008: 124-126).

In contrast, defensive pessimists typically set low expectations and prepare for various negative outcomes (Norem 2007: 121). Defensive pessimists create specific scenarios that could possibly go wrong and prepare against these outcomes (Norem 2007: 123-124). Surprisingly, however, defensive pessimists perform well and may even outperform strategic optimists in test situations, if their overall mood is *not positive* (Norem and Illingworth 2003: 352). The differences between strategic optimists and defensive pessimists are a learner difference, but they relate to language pedagogy in another

level. In the *Information Processing* paradigm of learning, input is central, and *mood is also input* (Norem and Illingworth 2003: 352-353). Therefore, engaging self-study materials should strive, if at all possible, to take mood into account.

In language pedagogy, there seems to bias towards positive approaches, perhaps arising from the dominant position of American linguistics in the field of English language pedagogy. There are sound reasons arising from psychological research for this positive bias, as the effect anxiety has on performance varies greatly between types of tasks. The effect on verbal fluency in speech is the greatest, whereas spatial reasoning remains intact (Bialystok 2002: 236-237). This is particularly the case for introverted learners, as they may suffer from a communicative breakdown, where the realisation that anxiety about their anxiety creates a downward spiral that prevents fluent speech (Bialystok 2002: 241). In the case of students who abhor the use of L2 and language learning, it is logical to presume that a recurrently negative mood associated with contexts of language learning or L2 use would then serve as input that 'scrambles' the linguistic input and hinders or prevents learning.

Methods and tools that utilize visual and spatial reasoning in language learning, such as word clouds, TermsBerry and GraphColl, have great potential in allowing students whose anxiety has a deleterious effect on their verbal performance to work with texts. Furthermore, successes in doing so may alleviate the students' anxiety, and in turn, help them to break out from the circle of negativity. Keeping the focus on developing competence in academic written communication, the utilized solution in the present study is to utilize the discrepancy in spatial performance and verbal performance, which in this case means the visualisation of texts. As noted in other parts of the thesis, there are other benefits to this, but the anxiety-reducing effect of allowing the student user to utilize spatial processing at least partially in some of the tasks might well have a psychologically positive effect in reducing the negative spiral of anxiety.

3.2.4 Colour, affect, and cognition: the benefit of colour for textual analysis

The effect various colours have on the mind has been a topic of speculation for centuries. The traditional division, which is still by and large in use is based Goethe's taxonomy of colours back in 1810, in *Theory of Colour* (Arnheim 2004: 358; Elliot 2015: 1). Newer psychological research has confirmed that colours are psychological experiences (Kurt and Kingsley Osueke 2014: 2), which has had implications for architecture and interior design of study spaces (Kurt and Kingsley Osueke 2014), even before there has been empirical evidence of the effect. So far, the effect of coloured text

on students' psychological well-being, to my knowledge, has not been researched, even though coloured texts are used frequently, as is highlighting of text in all educational levels. However, they are not used due to their possible mood-enhancing effect, but for other reasons, such as 'coding' different sections of a text. For moods, the basic principle: colours having an effect on the endocrine system via the hypothalamus in the brain (Kurt and Kingsley Osueke 2014: 4), by all logic, should remain the same: therefore, coloured texts should affect moods. Different colours have different effects on emotion and cognition. Previous research has shown that red undermines cognitively challenging texts, while blue facilitates alertness in attention-demanding tasks (Elliot 2015: 3). Colour also has behavioural effects, a fact well known in graphic design (Drew and Meyer 2008: 195). and increasingly in education, especially on the primary level and special education.

The frequent use of colours in these educational contexts has a basis on early research on perceptual behaviour – for example, Rorschach found out that depressive people generally react to shape while people with cheerful mood react to colour (Arnheim 2004: 335). Sensitivity to colour is linked to a *passively receptive* mind, while sensitivity to the shape of an object is linked to the *actively organizing* mind – due to the fact that the response elicited by colour is mainly affective while the response to shape is more intellectual (Arnheim 2004: 336). However, this is precisely the distinction between the superior cognition and inferior emotion that Vygotsky's work sought to challenge – and the increasing use of colour in the contexts of designing study space attests that the view of Arnheim – originally from 1957 from the first edition of his book *Art and Visual Perception* – may be falling out of favour, or at least it is being challenged. Nevertheless, it is widely agreed upon that colour has an emotional effect – based on outside observations on people and also on knowledge of neurophysiology. This effect both motor and glandular, as well conscious psychological: there is both a glandular effect that people are unaware of, and a – possibly learned cultural – response to certain colours (Drew and Meyer 2008: 196-197).

Based on the research on colours (Arnheim 2004; Kurt and Kingsley Osueke 2014; Elliot 2015), it can be assumed that students, particularly those whose state of mind and mood is not optimal, would benefit from the stimuli provided by coloured text when working with challenging texts. *Voyant Tools* provides the opportunity both a word cloud for frequency analysis that is at a sufficient level for the purposes of language learning, as well as a coloured multi-line graph that tracks the language over the course of the entire text. GraphColl, the other tool that was considered for the present study, also adds colour. This may have a beneficial psychological effect, but in it, it is mainly to clarify the

hierarchical relationship that the collocations have with each other, e.g. the original word that serves as a starting point, first-order collocations, second-order collocations, and so on.

In addition to the positive effect colour has on moods, colour in text and other forms of presentation serves the purpose of structuring and organizing the text. In contrast to mere differences in shape – for example, if Voyant Tools collocation view or GraphColl were to present the information in a way that node words and the collocates were all in black, it would be far less clear than it is: the colour option helps the user to better understand the network of collocations and keep track of the relationships the words have. This is because colour is the most efficient way of discrimination (Arnheim 2004: 330). Differences of *type*, are easier to tell apart from each other than differences of *degree*– this is the advantage a clear difference in colour has. For the human mind, differences of type are easier to keep in mind than differences of degree (Arnheim 2004: 333). Therefore, having the hierarchical relations between words distinguished by type by colour helps in separating the 'starting point' data from the linked data in the analysis of collocational networks.

A natural corollary to the relationship and benefits conferred by differences of shape and differences of colour, is that as shape variations contain nearly infinite amounts variation, and is, in the end, more distinctive means of identification than differences in colour, as it might come down to differences in hue (Arnheim 2004: 333). It is easier to conceptualize, while colour is difficult (Arnheim 2004: 332-333) – for example, it is easier to have an agreement about what kind of shapes *round* and *square* are than about variations in colour such as *light blue* vs. *teal*. A combination of clear differences in colour in sufficiently distinctively shaped presentation of data, which is something that the corpus tools utilized here have – should provide the maximal cognitive benefit from the viewpoint of being able to keep the information in mind. Of course, one obvious difference of shape is in typology – the most local level of factor of analysis in the analysis of texts (Halliday and Mathiessen 2014: 27; see Figure 1 in chapter 2.1.2 in this thesis).

3.3. Learner Variables

In this concluding part of the theory section of the thesis, I present a brief overview of two pedagogical considerations that are relevant to the target audience of the materials. First, the materials are mostly for adults utilizing them for EAP in conjunction with their field of study: it could be said that the materials are adaptable for *Content Language Integrated Learning*, CLIL. Of course, the phraseological focus, or textual focus (Nurmukhadekov 2015: 11-12) that comes with the teaching of

collocations and lexical bundles is more 'purely' language learning, but even then, the focus is on collocations in *academic* texts and lexical bundles in academic texts. Therefore, the learners are likely adults, and a review of the advantages and disadvantages adults have in language learning is in order. Or simply, the advantages and disadvantages adults have in learning in general – as the information processing paradigm (Saville-Troike and Barto 2017: 77-78) does not draw a distinction between language learning and other types of learning.

Second, a review of other factors, such as aptitude, motivation, cognitive style, and the different learning styles (Saville-Troike and Barto 2017: 90-99) that the users of the materials, the learners may have – learner types – concludes the theory section. Learner types is a subject that has been touched upon in other parts of the thesis, for example when reviewing corpus pedagogy in section 2.2., and in the review of cognitive, psychological, and affective factors affecting the learning process, section 3.3.2. I present the issue in a more general fashion that is still relevant to the context of the present, in addition to summarizing the previous sections.

3.3.1. Age and other demographic factors in corpus pedagogic EAP

Due to the superior brain plasticity that children and youngsters possess, the learning of new skills or languages is generally considered to be easier for them. This is particularly true in languages: the observation has led to the *critical period* hypothesis, the idea that language learning becomes progressively harder with age (Thiessen, Girard and Erickson 2016: 276-277). In L2 learning, the issue is more controversial. The benefit conferred by brain plasticity certainly benefits younger learners of L2, and current national-level recommendations stress the early, or at least earlier than the previous policies, learning of foreign or second domestic language. Despite these benefits and recommendations, which I do not contest, adult learners do possess certain advantages in language learning that younger learners often lack. These benefits make a systematic approach to language learning particularly suitable for them. Saville-Troike and Barto (2017: 88-90) break down the benefits that learners of different ages have in a table format with explanations, which I have adapted below from the viewpoint of corpus pedagogy by adding a column that reviews, which age group has a benefit using the materials presented here:

Table 4: Effect of age and **assumed** suitability of corpus linguistics pedagogy (based and adapted from Saville-Troike and Barto 2017:88).

Advantage: young	Advantage: old	Advantage young/old, or no assumption	Reasoning for advantage, or no advantage

(Saville-Troike and Barto 2017)	(Saville-Troike and Barto 2017)	(presumed)	(presumed)
Brain plasticity: easy to develop new neural pathways	Learning capacity: cognitive and *metacognitive* skills	Young	Learn to use the new tools faster. However, the advantage may switch in cognitively demanding tasks
Not analytical: less inhibitions and less self-monitoring when using language	More analytical: capable of breaking down language rules and learning grammar	Old	Better suited for *systematic* work with corpus tools
Less anxiety: anxiety hampers verbal reasoning (see 3.2.3)	Transfer from L1: helpful especially in related languages	-	-
Weaker group identity: less possible negative associations with a foreign language	Real World knowledge: can utilize previous general knowledge* in the process of learning about the language	Old	Particularly for compiling a corpus of one's own: have a personal goal linked to real World knowledge
Less concern for others: less inhibitions for *productive* language use	Pragmatics: understanding the contexts of language use better	Old	Corpus tool use does not include productive spoken language use much, and does include context
More intuitive, and potential to become native-like in the language	Better memory for vocabulary learning and grammatical rules	Old	Helpful for older students to address the lack of native-like intuition

*: A related issue is socio-indexical inferencing: learners can utilize their knowledge of socio-indexical categories (gender, class, etc.) in learning additional languages (Bosena, Fine, Kleinschmidt, Jaeger: 2016). This is an example of real world knowledge that older learners have where they might have an advantage.

As seen in the table above, it can be assumed that corpus pedagogy can be assumed to be especially helpful for older students. However, the notable exception is the ability to use new digital tools. The adroitness and enthusiasm commonly displayed by children and youngsters in utilizing new digital tools can be observed practically everywhere in the digital era where such tools can be found. Of course there are exceptions on both age groups in this, but the general advantage younger learners have here is clear. However, much of the advantage younger learners have can be explained by the predominance of *integrating* processes in favour of *extracting* processes (Thiessen et al. 2016: 279). That is, the younger learner builds connections from the received language input more readily, while the type of processes the older learner is suited for are *extracting*, or breaking the input into parts. In

addition, adult learners store, or encode, the input of new language data in a more precise manner – due to having a more mature memory system (Thiessen et al. 2016: 282). Generally, this is not an advantage in language learning since children and younger learners are capable of enriching and regularizing simple linguistic data into rules (ibid.). However, it might give adult learners an advantage in EAP using corpus tools, as it requires precision, rather than forming rules based on incomplete data.

Despite some stereotypical notions on the effect of physical sex and gender on the ability to learn a language. Specifically, it is often claimed that girls and women would possess an advantage in languages while boys and men would correspondingly have an advantage in maths. Empirical studies on the subject have found that such an effect does not exist, at least in language learning (Saville-Troike and Barto 2017: 90). Culturally ingrained notions on men and women may still be relevant when assessing the suitability of the tools presented here, but they were a minor consideration in the design of these tools. Namely, there was a rough presumption that the technical side of the tools might appeal to men more, while the visual, colourful aspect to women – but such an assumption is rather speculative. Other demographic factors, such as the possible ethnicity, or socioeconomic status of the users, are not relevant.

3.3.2. Mental factors in learning EAP

In this section, I review three mental factors affecting the language learning process: 1) aptitude, 2) motivation, 3) cognitive, metacognitive, and affective strategies (Saville-Troike and Barto 2017: 90-98). Of course, merely knowing about these factors may not be practically useful: to use Mokyr's (2005) classification, the theoretical omega-knowledge of these factors does not automatically turn into practical lambda-knowledge. Therefore, what follows is a review on what traits a 'good language learner' and how, and if, the language learning behaviours of a student can change or be modified (Saville-Troike and Barto 2017: 98-99). In particular, for the design of the material package, I am interested in how the corpus linguistic tools can make up for possible gaps in aptitude, be motivating, and work as a strategy to learn EAP. Therefore, each of the three aspects is reviewed, in relation to their relevance to the materials in the appendix. A fourth relevant factor is that of *cognitive style*. However, the topic is far too broad to be covered here, so they are covered cursorily, only in the table.

Even though an egalitarian attitude to learners is something that a teacher, or a designer of teaching materials should aspire to, it is a readily observable fact that learners do, in fact, vary in their

capability to learn. The ideal situation is that everyone learns – but not everyone is equally willing to extend the effort to do so, nor is everyone equally capable to do so. The capability to learn a language is referred to as language learning *aptitude*. For the sake of an argument, it is not necessary to consider whether language learning is the same as the learning of any other cognitive skill for now, as in the information processing viewpoint. Language learning aptitude is comprised of four cognitive factors: 1) phonemic coding ability, 2) the capability to learn language inductively, 3) sensitivity to grammar, and 4) associative memory (Saville-Troike and Barto 2017: 90-91).

Factors 2-4 are all of central importance in the present study, and sought to be taken into account. However, should the learner not have a high aptitude in all or some of them, it is to be hoped for that the tools provide additional support to aid them. In particular, collocational networks (Baker 2016) should provide assistance in the working of 4), associative memory: the associations within the text become visible and concrete and can be thus reviewed easily. For point 2), inductive learning, the process of using corpus tools should help to develop inductive thinking: the learner discovers things from the text with the help of *Voyant Tools*, for example. Grammar sensitivity, the third point is the least addressed here, at least explicitly, however, lexical bundles serve a lexicogrammatical function in the text and can be seen as part of grammar. In addition, one of the main uses of *concordance tables* in language learning is the discovery of correct prepositions to go with nouns – a rather grammatical task, done inductively.

Anyone who has ever tried to learn anything knows that the desire to learn – motivation – has a tremendous impact on learning. A mere desire to learn, however, does not yet equal to motivation, it is just one component of motivation. Rather, motivation is comprised, just like aptitude, of several distinct, yet interlinked components. To be motivated, the learner needs to have 1) a *goal*, a certain target. An example goal would be “to complete a thesis”. Second, obviously, is “) the *desire* to attain the said goal. Third, 3) the goal needs to be relevant to a need – it is hard to stay motivated if the goal cannot be seen as fulfilling a need. In the completing of a thesis example, the goal serves the need to graduate and get a job, for example. Other factors of motivation are ones that are not as intuitively understood as motivation. Fourth, the learner needs to have 4) *a belief in their success*, or perhaps be driven by a fear of failure as was the case with defensive pessimists (Norem 2007: 121). Fifth, a clear view of the 5) potential outcomes and rewards can be motivating, perhaps again affected, as appropriate by the variable of strategic optimism, where a positive outcome is visualized, or by defensive pessimism where a negative outcome is visualized and fought against. Sixth, the learners 6) self-concept, and their identity, is a factor in motivation. (Saville-Troike and Barto 2017: 91-92).

The last factor, self-concept and identity as a learner, refers to whether the learner sees themselves as someone who can, and wants to learn a new skill, e.g. “I am someone who can learn a new language” or “I am a hard-working student who does not give up easily”.

To make use of the aptitude, and to follow up on their motivation to learn, the learner needs to possess strategies to learn, either consciously or at least unconsciously. In the case of older students, as seen in table 3 in the previous section, the learners likely possess many or at least some strategies. As was the case with aptitude and motivation, strategies can also be divided into categories. Cognitive and metacognitive strategies, or skills, were already mentioned in the aforementioned table. In addition to them, *affective* strategies need to be considered. In the division presented by Saville-Troike and Barto (2017: 98) affective strategies are clearly different and distinct from the cognitive ones – a division that is contested by the viewpoint of Vygotsky, where cognition and affect are interlinked, and at least challenged in the information processing paradigm and ‘mood as input’ model in psychology (Norem and Ellingworth 2003: 352-353). Notwithstanding these controversies, the three categories naturally consist of a collection of strategies, with cognitive strategies and metacognitive strategies being most intrinsically linked together: as metacognition can be defined as ‘thinking about thinking’.

Cognitive strategies mainly consist of basic ‘school-type’ strategies for learning: repeating, translating, remembering, imagery, inferencing (Saville-Troike and Barto 2017: 98). The present study supports repeating, remembering and inferencing - and while the imagery used as a strategy refers to visual imagination rather than concrete visualisation of the text, the visual representation via word clouds, explicit collocational networks, concordance tables, and graphs may serve as a cognitive strategy. Metacognitive strategies refer to strategies that typically are learned later in life or the course of studies: previewing, attending to language input, rehearsing the components of the language, and self-monitoring of one’s own progress (ibid.). *Voyant Tools* can function as an additional tool for the strategy of rehearsing: previewing the main vocabulary with a word cloud of a text is a form of preview. The more specific, local-level analysis tools that come with *Voyant Tools* are about attending the input from the texts, and rehearsing the components – the particularly central components of a text, a task which also helps in guiding the central processing. Self-monitoring is a metacognitive strategy that typically takes some effort: in the form of reflection, for example. Self-monitoring is not solely a *good* strategy: excessive self-monitoring can hamper language learning, the learner becomes too conscious of the errors and the ‘areas of improvement’ to progress smoothly. Therefore, to be effective, it needs to be periodical and focused on the progress – not necessarily constant.

Affective strategies, in the context that Saville-Troike and Barto (2017: 98) present them, might be appropriately termed *social* strategies: most of them refer to some kind of involvement with another human being rather than strategies to regulate one's emotions in the learning process. Of course, the process of interacting with other human beings process might involve emotions. The strategies that are termed affective include interacting with native speakers of the language; working with peers; asking for clarification, presumably a teacher or a peer; and similarly asking for repetition and examples. Utilizing corpora for language learning can be done socially, in a way that involves the use of affective strategies, or in a teacher-directed manner that similarly presents an opportunity to ask for clarification. The interaction with native speakers is indirect and computer-mediated with corpora, and not real-time: the learners work with already existing texts. However, as mentioned in a different section, corpora, with the appropriate corpus tools, can function as "a tireless native-speaker informant" (Barnbrook 1996:140; cited in Römer 2011:214) to provide engagement with authentic examples of native use of the language, rather than text-book pedagogical texts made solely for learning purposes.

To summarize the mental factors presented above, I have adapted them into a table format below. Note that the table does not correspond horizontally, e.g. it should be viewed as a collection of the mental factors involved in language learning. Strategies are divided by cognitive (C), metacognitive (M), and affective (A). The cognitive styles are generally contrasted in a binary way: in the table they are presented in the form of the contrasted style A and B – with no value judgement on the better or worse style. In the table, the relevant aspects or strategies are shaded light orange: they are aspects that are in one way or the other covered by in the materials in the appendix of the present study.

Table 5: A summary of categories of Aptitude, Motivation, and Strategies + a brief overview of the contrasted Cognitive styles. Based on Saville-Troike and Barto (2017: 90-98).

Aptitude	Motivation	Strategies: (C/M/A)	Cognitive style (Contrast A)	Cognitive style (Contrast B)
Phonemic coding	Goal	Repeating (C)	Global	Particular (or local)
Inductive capability	Desire	Remembering (C)	Deductive	Inductive
Grammar sensitivity	Relevance of goal	Rehearsing (C)	Focus on meaning	Focus on form
Associative memory	Belief in success*	Inferencing (C)	Holistic	Analytical
	View of outcomes	Imagery (C)	Field-dependent	Field-independent
	Self-concept	Previewing (M)		
		Attending to input (M)		
		Rehearsing components (M)		
		Self-monitoring progress (M)		
		Interacting with native speakers (A)		
		Working with peers (A)		
		Asking for clarification (A)		

*: Alternatively, a defensive pessimist strategy to prepare well against negative outcomes (Norem and Illingworth 2003; Norem 2008)

Finally, there is the question of what makes a good learner, and whether this good learning behaviour can be initiated or changed or not. The last question is yet unknown. Saville-Troike and Barto (2017: 98-99) summarize the findings of research on the subject and claim that good learners are characterized by the following traits:

1. Concern for form,
2. Concern for communication,
3. An active task approach,
4. Awareness of the process,
5. Capacity to use strategies (outlined above),
6. and flexibility in accordance to the requirements.

The finding that good language learners specifically pay attention to form suggests that an analytical, corpus-based approach that pays explicit attention to form is efficient. It pays attention to typographical, orthographical and lexicogrammatical features in addition to addressing the communicative purposes of the texts, their socio-semiotic processes. It is an approach that utilizes and strengthen the two first traits. Hopefully the corpus tools tasks presented in the appendix will further an active task approach, awareness, and increase the capacity to use strategies.

4. The material package: corpus programs for learning EAP

In this chapter, I present the materials in the appendix, and explain what theoretical principles they are based on. It consists of three parts. First, an overview of the materials and what is the reasoning behind the whole package. Second, an overview of the computer tools used in them and a comparison of them. Third, I present the reasoning behind each individual task in the appendix, broken down in groups that make pedagogical sense. The ordering is from easier to harder, for the purpose of scaffolding the learning process. Since the learner has to not only learn EAP but the use of new technical tools, it is imperative to not overload or discourage the user. Thus, every time a new tool is introduced, the linguistic challenge level drops, as it is necessary to learn the new tool. This way the learner or user can also choose to learn just one or two of the tools and does not have to go through the tutorial parts of the others. Thus, the material package, fittingly for one based on corpus linguistics, is modular and broken down into parts.

4.1 Overview of the materials

The materials created on the basis of the theory presented in the earlier chapter are designed to function either as a self-study package, or as something that a teacher of Academic English can include in his or her course. The first option is the default: the conversion tips are offered as part of the material package in the appendix. The level of the materials ranges from junior high school level

for the simpler tasks utilizing word clouds, to university / polytechnic level. The default user is a university student. Of course, it is entirely possible that there are younger students who can handle the more difficult materials, and previous research suggests that not all university level students will continue using corpus tools even after they have attended a course where they compile and use a corpus to assist them in writing tasks (Charles 2014). Similarly, it is likely that not all intended users will find the materials appealing for various reasons. Still, it is my hope that learners and teachers can find at least something to use, with the dictum of "what can I use" (Kennedy and Miceli 2017: 111).

The material package, and the ordering of tasks in it takes constructivist principles and the cognitive framework of information processing into account. Because knowledge is constructed by the learner in constructivism based on probable truths (Tourmen 2016: 9), the users will be working with texts of their choice, texts that are relevant to their own discipline and study or research interests. Second, because of the importance of input for intake, the learning of language, within the information processing framework (Saville-Troike and Barto 2017: 81), it is necessary to have a gradual, continuous input of new information. Thus, the tasks progress from easier to harder. This is to reduce the cognitive load of learning too many things at once: the first tasks are far less challenging from the viewpoint of language learning, as the user will be familiarizing him/herself with the corpus tools and the basic logic of corpus linguistics. This familiarization includes gaining familiarity with the basic terms of corpus linguistics. For this purpose, a glossary of the terminology is presented in the beginning the terms are explicated in the tasks whenever possible. However, since memorizing glossary before being able to work on to the tasks would be inconvenient, the terms are either explicated or 'plain English' is used in the instructions for the user, and the glossary is something that the learner can go back to.

The results of the previous studies (Lee and Swales 2006; Charles 2014) were utilized in the design of the materials to make them more widely appealing. The concern of inconvenience and difficulty of use, factors mentioned by the participants of the corpus linguistic writing course who discontinued their use of personal corpora (Charles 2014: 34-36), were considered. Notably, at the time of the aforementioned previous studies, especially the first one, there were not as many corpus programs available as there are today. For example, the main program utilized in the present study, *Voyant Tools*, has been available only since 2015. The self-study package is built mostly around it, because it can function as a pedagogical tool for learners and teachers alike. However, for the study of concordances, a central pedagogic use of corpora, *AntConc* remains the superior choice. Therefore, tasks, where the users are taught to build their personal corpora also for the analysis on *AntConc* are

included. They are tasks that come later: the reasoning is to not overload or discourage the user from corpus tools, but to slowly scaffold the learning. In a similar manner to the ordering of the tasks where the user works with *Voyant Tools*, the idea is to first gain familiarity with *AntConc* before using it for more linguistically challenging tasks: the discovery of lexical bundles and for analysing the immediate textual context of a central term from the personal corpus. In other words, basics first.

4.2. Review and comparison of the three corpus programs

The materials presented in the index are meant to be used with a computer: the paper, or text file version contains screenshots from the relevant programs and step-by-step instructions on using them. To reiterate, the programs used in the materials are *Voyant Tools, AntConc*, and *Graphcoll*. In this chapter, I review these three programs as learning tools for corpus pedagogy first individually, and then compare them. This comparison is for the benefit of the users: some of the tools are better suited for certain types of analysis, others are easier and thus offer a lower threshold entry to corpus analysis.

4.2.1. Voyant Tools

Voyant Tools is a fairly new, online, browser-based corpus analysis tool based on the research by Geoffrey Rockwell and Stephen Sinclair, presented in their book *Hermeneutica: Computer-assisted interpretation in the humanities* (2016). It is an updated and revised version of an earlier program of an earlier program known as *Voyeur*. The toolset and the book were a part of larger project, hermeneutica.ca. (Graham, Milligan, and Weingart 2013: 36). By March 2019, it is linked to two larger corpora by default. Both are specialized: a corpus of Shakespeare's works and a corpus of Jane Austen's novels. As it is still being updated, more corpora might be added: most likely for the analysis of literature. Due to its origins, Voyant Tools is designed for use in the Humanities, so users from other disciplines will not likely find much use from the default options of Shakespeare and Jane Austen. Users interested in analysing other types of texts have to create their own corpora to analyse. Fortunately, creating a personal corpora is rather simple with Voyant Tools: the fourth task in the materials linked in the appendix gives step-by-step instructions. These instructions should be clear enough for users who do not find the use of the program intuitive, for some reason.

The opening page of Voyant Tools is simple, appealing, and clear: there is a text box in the middle and an option of opening one of the two corpora, uploading a file, or entering a text into the box in the middle by example by copy-pasting it from somewhere else. The simple opening page belies the possibilities offered by the analysis page. In contrast, a plethora of information is presented in the

next page. This amount is likely discouraging to users with no familiarity of the terminology presented in it: the only element that is likely to be familiar to the user is the word cloud on the upper left.

4.2.2. AntConc

AntConc is a downloadable freeware program dedicated to the analysis of concordance tables. It is available on multiple operating systems: Apple OS, Windows, and Linux. AntConc requires installing, which means that the user needs administrator rights or at least extended rights, for the installation. There are upsides and downsides to this: when the program has been downloaded and installed it can be accessed reliably from the user's computer. There is no need for an internet connection. The downside is that the program is not necessarily available on a public computer that a user is working with. With AntConc, the user can either study their own corpus or corpora, or access the frequency lists of two extensive general corpora of English texts: British National Corpus, BNC, frequency list for BrE; and Brown's national corpus frequency list for AmE.

A dedicated concordance program, AntConc has been successfully used for the teaching of EAP in the past (Charles 2014). Therefore, it can be said that AntConc has an established position in EAP corpus pedagogy utilizing concordances. The program itself is fairly user friendly and can the principles can be grasped quickly: the locating of collocates from the concordance table is particularly easy with AntConc. It has buttons that specify the distance and direction of any of the four collocates that can be looked for simultaneously. The collocates are presented in different colour from the other words in the table of concordances.

In contrast to the fact that AntConc is supported on most operating systems, AntConc is limited in its ability to read different file formats. This presents at least a minor obstacle to the user. Most downloadable academic articles are in .pdf format. This means that the user has to convert these texts to .txt format for the analysis of a self-compiled corpus of academic texts. The same problem exists with the creation of a corpus of learner texts. These texts are typically in Ms Word file format, .docx. This shortcoming can be utilized to make the corpus created with AntConc better, similar to how corpus-building has involved a cleaning up phase in a previous study (Charles 2014: 31). The conversion should be accompanied by a cleaning up, removing numerals and non-words from the text files and combining them together, into one long text. Further, .txt files are extremely small bit-wise by today's standards so these can be easily transported by a memory stick, stored in a cloud, or even kept as an email to oneself, so it can be accessed in the inbox in every place with an internet access.

4.2.3. GraphColl on #Lancsbox

GraphColl, which nowadays is a part of larger corpus toolkit called #Lancsbox, is a program that needs to be downloaded for use from its home page, which is affiliated with Lancaster University. Multiple operating systems are supported: Windows, Apple OS, and Linux. Currently, it is only for analysing texts in English. Graphcoll, when it used was separate from #Lancsbox, was a *portable software*, the user supposedly only needs the files downloaded on a computer or even an USB memory stick in order to be able to use it and continue using it. An internet connection is not needed after it has been downloaded. However, most public computers do not permit guest users to execute the .bat file, which is a command prompt file, required to run the program for reasons of data security. Nowadays, the entire program, #Lancsbox, is an installable program like AntConc, a fact that practically makes no difference anyway.

GraphColl requires some effort from the user in learning the logic of the algorithms it uses to create the tables and the presentation of the collocation network. Mainly, the user has to be able to understand the cut-off point for the *strength of association* between the words included in the graph. If the cut-off point is set as too high, e.g. the definition of collocation is too strict, the graph will present too little information, it is unable to find collocates. If it is set too low, the graph will be extremely cluttered, making analysis impossible. Fortunately, a helpful video tutorial is available in the website. In addition, the materials presented in the present study include step-by-step instructions with screenshots with GraphColl, as was the case with Voaynt Tools and AntConc.

Unfortunately, despite the earlier intention to do so, #LancsBox was not utilized in the materials of the present study. The program, when installed, performed inadequately and was unable to load Brown's corpus for analysis. However, the use of two corpus programs with different foci is already plenty for a Master's level study, so the study, and the user of the materials, can survive without them.

4.2.4. Comparison of the three programs

In this chapter, I present a table of the advantages and disadvantages of the three programs. First, by including technical considerations, visual appeal and convenience, and the types of users they are suited for. Second, and more importantly, the table contains a comparison of the usability of the three programs on analysing different aspects that are typically studies in corpus linguistics: frequencies, collocations, concordance tables, and lexical bundles. Third, the programs are compared on how well they suit the pedagogical goals of this thesis: how well they are suited for analysing vocabulary and

lexicogrammatical features such as lexical bundles; and how well they are suited for analysing concepts and discourse features related to them, e.g. how a certain term is used in the context of the article or articles it appears in.

Table 6: Comparison of three corpus tools meant for use in the present study.

Trait ↓ Tool →	Voyant Tools	AntConc	(GraphColl)
Visual appeal	+++	+	++
Visualisation of data	+++	+	+++
Portability	+++	+	++(+)
Ease of use	++	+	+
Frequencies	+++	+	+
Collocations	++	+++	+++
Concordance table	+	+++	++
Lexical Bundles	++	+	+
Linguistic reference	++	+++	++
Concept analysis	++	+	++
Discourse links	++	+	++
Analysis of general corpora	+	++	++
Analysis of personal gathered corpora	++	++	++
Analysis of learner corpora	++	++	+

+: Can be used for this purpose ++: Well suited for the purpose +++: Very well suited for the purpose

Note that none of the tools are exceptionally well suited for learning about lexical bundles. The best options are offered by Voyant Tools, but there is considerable work to be done in separating lexical bundles from other phrases occurring in the article. The grouping of all frequent phrases together is

not necessarily a bad thing, but it means that the theoretical goal of separating phrases, idioms, and multi-lexeme terms from lexical bundles is not plausible. Consequently, the lexical bundle focus in the present study is not as extensive as the focus on frequencies, collocations, and collocation networks. Concept analysis is also something that none of the tools is dedicated to by itself, but the focus on these is still strong in the materials, because of their importance in EAP. Indeed, the focus is on making the tools work for this purpose in EAP.

4.3. The individual tasks of the material package

In this section, I present the rationale behind each of the tasks. Due to the amount and size difference between the tasks, they are presented in groups. These groups are related to pedagogical goals of mastering the use of the tool; grasping the principles of corpus linguistics; and learning language either on the level of vocabulary, or discourse. In the beginning of each group of tasks dedicated to a tool, the learning of these tools is the primary goal. The learning of the principles of corpus linguistics is something that occurs throughout.

All of the tasks follow a similar structure, a formula. First, the number of the task and the name. Second, an explanation of what the user will do and the purpose of the task. Third, the learning objectives in a bullet point format: for quick reference about the benefits offered by grasping the skills taught in the task. Fourth, the instructions. The instructions part is, by far, the largest: it contains step by step instructions on what is done at the task, interspersed with numerous screen shots. The idea is to make the tasks as technically simple as possible for the users, taking into consideration the fact that the greatest benefit offered by corpus linguistics is in the field of Humanities and Social Sciences, where these technical skills are not necessarily at the highest level.

4.3.1. Tasks 1-3: Word clouds and introduction to Voyant Tools

These three tasks introduce the user to *Voyant Tools* and the most visible application of corpus linguistics: the ubiquitous word cloud. The first task is based on the pedagogical use of word clouds by Olga Filatova (2016). In it, the user learns a new strategy for L2 texts: the previewing of the most frequent vocabulary in the form of a word cloud. It is a metacognitive strategy, since preview strategies are classified as metacognitive strategies (Saville-Troike and Barto 2017: 98). This task is well suited to other educational context apart from university EAP learning. It is a strategy that can be utilized even at level of elementary education with teacher guidance. Whether this method of previewing articles is something that the users of the materials will continue using is hard to predict,

but it is likely that it is too simple for the target user group. For that reason, the task is to introduce the user to *Voyant Tools*. The instructions explicitly state that more difficult tasks are to follow, in case the users get frustrated by the seeming simplicity of the first task.

The second task follows the first one closely in structure. Preferably, it should be done in right after doing the first one, as the user is expected to do a rough comparison of their own writing, as shown on the word cloud, with that of a professional writer. The targeted area of improvement is productive vocabulary: the goal is to make the help the user to notice if they overuse words and add diversity to their writing. The task is suited for students at an early stage of their university studies, or for younger students as long as the first task is similarly adapted. It is also suited for L2 writing pedagogy that is not about EAP. Any genre of writing is suited for task two paired with the first one. The second task also offers an additional benefit. The *Cirrus* feature of *Voyant Tools* becomes even more familiar to the user. At the end of the second task at minimum, the user should be comfortable with the use of *Voyant Tools* for word clouds.

Finally, in the first series of tasks, the third task primes the user for the tasks that follow it. In it, the user gains a first touch to the analysis of concepts with corpus linguistics and *Voyant Tools*. It is similar to the first one in the sense that it teaches a metacognitive preview strategy (Saville-Troike and Barto 2017: 98, see Table 5). It also utilizes the information processing principles of input and intake. It does this by teaching the user to attend to certain input more. Attending to input is also a metacognitive strategy (Saville-Troike and Barto 2017: 98). In it, the user is taught to use several options in *Voyant Tools* to identify the central, most frequently occurring concepts or terms in an article and pay closer attention to them. Thus, the user might learn to read articles for the most relevant information, and to perhaps even the option of skimming academic papers to see if they are relevant to their interests, an option that is not possible for all readers or with all texts (Biber and Conrad 2009: 111), if they do not already possess this strategy, or possess the inefficient version that hardly can be described as a strategy: reading carelessly. In addition, the task includes some writing: the writing of definitions of the concepts presented in an article. With this, the input and intake have chance of turning into *indwelling* of knowledge (Polanyi 1968; cited in Pyrko et al. 2017: 311). That is, by having to describe the concept in their own words, the knowledge becomes personal.

4.3.2. Tasks 4-6: Creating a mini-corpus with Voyant Tools and analysing it

These tasks introduce the user to what is perhaps the most pedagogically beneficial use of corpora in EAP: gathering a personal corpora and using it as a linguistic reference tool (Lee and Swales 2006; Charles 2014). Task four is a guided instruction for creating a corpus of research texts using *Voyant Tools*. Later on, the ability to build corpora learned in the task will be put to use for the creation of a larger reference corpus. This task is straightforward and pedagogically focused on building technical skills. The reason why it has no specific vocabulary or EAP related goals is that the technical skill learned in it is particularly important in relation to the larger pedagogical goals of the material package: compiling and utilizing a personal corpus for use as a linguistic reference, for studying the type of EAP in each learners' discipline of choice, and for learning about how concepts central to the discipline are addressed in a selection of texts from a field. Therefore, it is imperative for the learners to know how to build corpora, and also to have a task that serves as instructions to it should they forget how to do it, without mixing other pedagogical goals into it. Task four is marked as an *instructional* in the package. This way, it is clearly separated from the others.

In task five, the user learns how to use three different *Voyant Tools* options for finding differences in the definition of terms and concepts in different articles or academic texts. It is still based on frequency analysis, with the main difference from word clouds being that now the user learns how to track the occurrence of a term across articles and chronologically inside the text of an article, that is, whether the term is frequently throughout the text or that it occurs more in the beginning, the middle, the end, or otherwise irregularly in the text. Task five is similar to task three in the first 'series' as it is about the analysis of terms and concepts. Since the user understands the principles of term analysis now, from task three, has a corpus to work on from task four, the task can be more complex and challenging. Or, to put it in general terms, the user is familiar with using corpus linguistics as a concept analysis method and *Voyant Tools* as a program. The challenge should not be excessive, as the higher challenge level is something the users can now handle due to being *scaffolded* by the previous tasks.

Task six is again a technically challenging task, but it also introduces the user to central corpus linguistic concepts: collocation (Gries 2013; Brezina et al. 2015), and collocation networks (Baker 2016). It concludes the series of tasks working with the mini-corpus, and after completing this task, the user knows how to use all the visual tools in Voyant Tools. The task is long and multi-staged.

Fortunately, as uploading the corpus the user has made in task four is easy, they can go back to it easily should time constraints force the user to stop.

4.3.3. Tasks 7-9: Cleaning up texts and using a personal corpus for reference

In this series of tasks, the user is asked to create a reference corpus of at least twenty articles in their own field. As it concludes the work on Voyant Tools, this series of tasks is the most intensive one so far. Most of this intensity comes from task 7, where the user is introduced to the cleaning up process required to build a professional corpus. However, the cleaning process is advisable, even if it is partial: the corpus is much clearer and turns up more reliable data if it is at least partially cleaned (Charles 2014: 21).

In a pedagogic sense, task seven does not require or teach EAP except incidentally. Rather, it is mostly manual, strenuous work. Therefore, at least a partial cleaning should be encouraged, as there is a danger of the perceived amount of work becoming an obstacle to corpus building (Charles 2014: 36). Nevertheless, the task provides instructions for complete cleaning, with caveats that should the amount of work be too excessive, partial is better than none. This is also part of the rationale between the entire present study focusing on corpus tools: to lower the threshold on their use. After the user has completed task seven, they have a personal corpus that can be utilized in all the tasks to follow.

Task eight is more of a tutorial on how to use Voyant Tools as personal reference in the future than a task with a distinct pedagogic goal. In a sense, this series of tasks is to build independent use of corpora for the user, and give ideas on how to use Voyant Tools in their future career or studies. In it, the user learns to use all the rest of the options / boxes on the toolkit. Namely, the concordance table and phrases, combined with the ability to see the original source text in the reader window of Voyant Tools. This practice is to prepare for the last task in the series, the last task using Voyant Tools, the ninth task.

The ninth task is an essay writing task, and therefore teaches English for Academic Purposes the most. In it, the user is asked to write an essay about concepts in their field utilizing Voyant Tools as reference. In it, the interlinkedness of Academic English and Academic thinking (Nagy and Townsend 2012) becomes apparent: the words denoting concepts, and understanding how they are covered by different authors is also a useful research skill. The lack of hands-on guidance in it is deliberate. So far, the user has been guided through every step. Now, the user already knows of all

Voyant Tools options, so guidance is not necessary. It is explicitly stated that this task is particularly suitable for users who study in the Humanities or Social Sciences, as these are concept-heavy and language-heavy fields of study, where corpus tools have a natural place.

4.3.4. Tasks 10-12: Learning to use AntConc

This series of tasks teaches the user, who by now is familiar with the principles of corpus linguistics, and is effectively a corpus linguist, the use of AntConc. As a dedicated, long-standing, widely utilized corpus program, AntConc has some benefits over Voyant Tools that the user may be interested in. They are organized as a standalone package apart from task 7, the cleaning up task, which is also useful and needed here if the user wants to use AntConc as a linguistic reference.

It is possible that the user is discouraged from starting a new series of tasks after finishing the ones with Voyant Tools. Hopefully, this does not occur, as these tasks are particularly useful for the user in the future as they teach the use of corpora as linguistic reference even more. To reduce the threshold of once again starting to learn the skills required to use a new tool, task ten is once again a lighter one, a tutorial or introduction to AntConc. Because AntConc uses more corpus terminology in its terminology, one of the stated aims is learning corpus linguistic terminology with it.

After the tutorial, where the user has learned to discover frequencies with AntConc, use concordance tables in it, and discover lexical bundles with it, the user is asked to use these skills in the next task, the eleventh one. In it, the user studies their personal in order to see how words are used in context and what kinds of lexical bundles there are in their field of studies. The user is asked to write sentences utilizing their knowledge of the previously unfamiliar or partially familiar words and combine them with the lexical bundles. This means that the EAP focus is extremely clear here.

Finally, task twelve is again a contrast and compare task that pairs well with task eleven. In a sense, it encapsulates the end part of the title of the present study well: it is the *and beyond* to the *Word clouds* in task one and task two. In it, the user is asked to upload their own essay writing to AntConc so they can compare their use of words and lexical bundles with that of professional writers and notice areas of improvement. After completing it, the user has all the tools they need to use corpora to develop as writers of EAP based on their own goals and needs.

4.4. Adapting the materials

In this section, I present ideas on how to make the materials, or some of the materials work in other educational contexts than University EAP for L2 English users. The materials are intended to be used as a stand-alone self-study package, but using them in teacher-led contexts teaching EAP may uncover more possibilities, and be even more efficient. Therefore, the sub-sections of this chapter covers two themes. First, adapting the materials to be used by teachers in indirect and direct corpus pedagogy. Second, adapting the materials to other educational contexts, such as secondary education or vocational ESP learning. These adaptation ideas are interlinked. The materials are matched to the level of tertiary education, and with minor adaptations could be used by L2 learners with advanced English skills and familiarity of the content knowledge in their own field of studies. But, they might be too challenging for students whose L2 skills and academic thinking are not at that level, at least without support.

However, the presence of a teacher provides additional support, a zone of proximal development where the difficulty is above what the learner can, at that stage of learning, do by themselves, but can do with the help of a teacher. In this way, the maximum amount of learning can occur. Later on, the presence of a teacher is not necessary and the learner has the ability to accomplish more on their own. Further, using the materials in a more communal way advances a theoretical goal that academic language, and academic knowledge, exist in its *community of practice*: an academician, and indeed any writer, is involved with an audience, a community. Therefore, the opportunity for teacher and peer feedback is a beneficial addition, if possible.

4.4.1. Teacher-centred use of the material package

When considering teachers or teacher trainees as possible users of the material package, the most obvious use is indirect. One pedagogical use of corpus linguistics is they can be utilized indirectly by teachers to find out the most important language forms and terms to teach (Römer 2011). This means that teachers wishing to compile their own materials or make supplementary materials in any contexts can use the material package to learn additional skill. Teachers can use the materials in order to learn about utilizing general or self-made corpora to derive linguistic examples from for language teaching, and to acquire the concrete skills to do so.

In addition to indirect use, the self-study materials are adaptable to classroom or guided course use. In fact, there are advantages for doing so. The more technically savvy students can provide assistance

to the ones who find the tools more difficult to use, and students can of course provide linguistic assistance to each other. Students within the same field of study can collaboratively gather a larger, more varied corpus than just one student. As a specialized linguistic reference resource, a large corpus is more reliable than a smaller one. This is because the larger one is more representative sample of the writings in a field. There are more authors and more articles. And the insufficient or idiosyncratic article searching habits by one student do not affect the outcome, the entire corpus, much.

In teacher-centred use, the main advantage is the opportunity for pair work and group work: the students can become corpus researchers in groups rather than just by themselves. The teacher can also monitor the progress of the students, either in the classroom, or via assignments that students do related to the tasks. The teacher can, for example, increase or decrease the amount of work required in subsequent tasks based on their observations. For example, many of the tasks have a specified amount of words that the user is asked to choose for analysis, or a small written assignment – the amount of words or the length of the writing assignment can be adjusted. A teacher can also reorder some of the tasks, or include only some, as the material package is modular. For example, a teacher wishing to teach the use of concordance tables for collocations can choose the relevant tasks to teach: an AntConc tutorial and tasks related to it. Teachers wishing to utilize corpus linguistics outside EAP may utilize the materials. This is the topic of the next chapter.

4.4.2. The material package in other educational contexts or outside EAP

For the dual theoretical purpose of the present study – the teaching of academic language *and* academic thinking (Nagy and Townsend 2012), the intended target group of the material package was restricted to university students learning English as a second language. Therefore, users outside this context might not find the materials as optimally useful. But, since corpus linguistics, in itself, can be used to study all manner of texts, and even other types of communication, there is no reason why tasks that utilize corpus linguistics could not be adapted. They can be adapted to either other educational contexts as covered before, or to other varieties and genres of written English than EAP. The theoretical background sections of the present study provide a sound basis for conversion ideas, even if the focus in the analysis is the EAP and the discipline variations within it. Consequently, the materials include a section of conversion tips, which work best when the materials are also used as teacher materials, as the teacher can do the conversion of the tasks based on their pedagogical knowledge, and familiarity with the learners.

The simpler conversion involves changing the target group. Obviously, the less of a distance in educational level and context the conversion involves, the simpler the conversion. In addition, some of the more advanced tasks cannot be adapted to other contexts, at least contexts that are far away from the original. For example, compiling a large linguistic reference corpus is something that is probably unsuitable to at least early stages of secondary education, and most certainly unsuitable for primary education. On the other hand, some of the simpler tasks, such as word clouds, might work very well in primary education context or in special education: Filatova's (2016) experience in using word clouds suggests that they are particularly suitable to visual and kinaesthetic learners, who might struggle in education that is auditory in focus.

The harder conversion involves changing the target language variety, and the genre. The tasks are written with EAP in mind. In order for them to work for other language varieties, the instructions must be different. Nevertheless, some language varieties and genres are closer to academic writing, and the conversion can be done without an excessive amount of work. For example, newspaper articles are an interesting genre to look at with the help of corpus tools, and as a genre, newspaper articles share many features with academic articles (Biber and Conrad 2009: 111). In particular, the analysis of discourses surrounding controversial topics in newspaper corpora is a field of study where corpora have been utilized successfully (Baker 2016). This is a conversion that can be combined with the academic context: a researcher utilizing, or specializing, in discourse analysis, or the study of rhetoric, may utilize corpus analysis of a collection of newspaper texts as part of their research.

Voyant Tools and Graphcoll offer particularly good options for this sort of study, as they can be classified as theme-oriented corpus visualisation tools. Voyant Tools, in addition, has great options for document-oriented analysis (see table 2 on classification of text visualisation methods used in the present study). In other words, GraphColl and Voyant Tools can be used to uncover larger themes and discourses from the newspaper texts, not unlike how Paul Baker (2016) used GraphColl to study newspaper discourses and attitudes expressed in the articles related to British troops in Afghanistan. Voyant Tools can also distinguish between documents, which means that the user studying a large newspaper corpora can also distinguish in which documents, and to what extent, do certain terms and certain collocations occur in. For example, when analysing ideologically charged topics the user can then distinguish between categories: whether the newspaper article is, for example, against a military intervention or for it. This is a particularly fruitful opportunity for critical studies uncovering loaded language. Apart from some conversion ideas, and the fact that the tools can be used for all these, it is a topic best left for further studies and to the discretion and ingenuity of the users.

5. Discussion and Future

The present study is a contribution to the advancement of corpus methodology and principles in the learning of EAP. The purpose was to provide a low-threshold entry to the world of corpus linguistics and help the users of the materials develop in EAP and in their academic thinking. Although the idea at an early stage was not to introduce any theory or terminology to the user, the materials in the appendix turned out to be rather theoretically intensive. In retrospect, this is not a flaw, as the theoretical part of the present study repeatedly stressed the importance of correct, exact terminology in EAP. Therefore, materials that come up with new, lighter terminology for the important concepts in corpus linguistics that the user nevertheless will be utilizing would have been in disagreement with the very theoretical principles that it supposedly relies on.

Constructing the material package was, at times, a highly challenging. This was because the original idea – the use of word clouds in EAP, turned out be insufficient in teaching EAP beyond the level of single words, at the level of collocations, collocation networks, and lexical bundles. Word clouds have their limitations compared to other tools such as concordance tables, so most of the tasks in the materials – nine out of twelve – utilize something else. Nevertheless, word clouds serve in the function that they serve well in other contexts: as an introduction.

Hopefully, the users of the materials find them useful and develop new ideas in using corpus tools for their own studies. There was no opportunity to test their use on a group of users, so there is no empirical evidence on whether on how appealing the materials turn out to be for users. Based on earlier studies where corpus methodology was used to teach EAP (Lee and Swales 2006; Charles 2014), corpus methods could still be more appealing. The present study is a practical contribution to this observed gap in corpus pedagogy. It also seeks to advance the direct use of corpora in language learning, an even more underutilized resource (2017). More empirical studies of different users of corpora to learn EAP are needed.

Finally, even if corpus linguistics remain only a tangential interest to the users, it is also to at least know of a new method to analyze texts and discourse. The current tools offer a plethora of opportunities for learners, teachers, and researchers. The process of studying the use of tools and corpus linguistics as a method is a gold mine of research ideas in discourse studies far beyond the goals and the scope of the present study.

6. Bibliography

Anthony, L. (2019). AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Arnheim, R. (2004). *Art and visual perception: a psychology of the creative eye*. Oakland: CA: California University Press. Revised and expanded edition.

Baker, P. (2006). *Using Corpora in Discourse Analysis.* London: Continuum.

Baker, P. (2010). *Sociolinguistics and Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Baker, P. (2016). "The shapes of collocation." In *International Journal of Corpus Linguistics* 21 (2). 139-164.

Ballance, O.J. (2017). "Pedagogical models of concordance use: correlations between concordance user preferences." *Computer Assisted Language Learning.* 30 (3), 259-283.

Barnbrook, G., Zyngier, S. and Vander, V. (eds.) (2011). *Perspectives on Corpus Linguistics.* Amsterdam: John Benjamins Publishing Company.

Biber, D. and Conrad, S. (2001). "Quantitative Corpus-Based Research: Much More Than Bean Counting." *TESOL Quarterly* 35 (2). 331-336.

Biber, D., Conrad, S. and Cortes, V. (2004). "If you look at…: Lexical Bundles in University Teaching and Textbooks." *Applied Linguistics* 25 (3). 371-405.

Biber, D. (2006). *University Language: a corpus-based study of spoken and written registers.* Amsterdam: John Benjamins.

Biber, D. and Barbieri, F. (2007). "Lexical bundles in university spoken and written registers." *English for Specific Purposes* 26. 263-286.

Biber, D. and Conrad, S. (2009) *Register, genre, and style.* Cambridge: Cambridge University Press.

Biber, D. and Gray, B. (2016). *Grammatical Complexity in Academic English.* Cambridge: Cambridge University Press.

Brezina, V., McEnery, T. and Wattam, S. (2015). "Collocations in context: A new perspective on collocation networks." *International Journal of Corpus Linguistics* 20 (2). 139-173.

Brezina, V. (2017). *Statistics for Corpus Linguistics: Introduction.* Video lecture on September 12, 2017. http://corpora.lancs.ac.uk/lancsbox/materials.php.

Boulton, A. (2017). "Corpora in language teaching and learning." *Language Teaching* 50 (4). 483-506. doi:10.1017/S0261444817000167.

Bosena, P., Fine, A., Kleinschmidt, D., and Jaeger, F. (2016). "Learning Additional Languages as Hierarchical Probabilistic Inference: Insights from First Language Processing". *Language Learning* 66 (4): 900-944.

Charles, M. (2014). "Getting the Corpus Habit: EAP Students' long-term use of personal corpora". *English for Specific Purposes* 35 (2014). 30-40.

Clapp, M., Aurora, N., Herrera, L., Bhatia, M., Wilen, E. and Wakefield, S. (2017). "Gut microbiota's effect on mental health: the gut-brain axis." *Clinics and Practice* volume 7:987, 131-136.

Coccetta, F. (2011). Multimodal functional-notional concordancing. In Frankenberg-Garcia, A., Flowerdew, L. and Aston, G. (Eds.) *New Trends in Corpora and Language Learning.* London: Continuum. 121-138.

Cogo, A. and Dewey, M. (2012). *Analysing English as a Lingua Franca: a Corpus-driven Investigation.* Continuum International Publishing.

Coxhead, A. (2000). "A New Academic Word List." *TESOL Quarterly*, 34(2): 213-238.

Coxhead, A. (2012). "Academic Vocabulary, Writing and English for Academic Purposes: Perspectives from Second Language Learners". *RELC Journal* 43 (1), 137-145.

Drew, J. and Meyer, S. (2008). *Color Management: A Comprehensive Guide for Graphic Designers*. Mies, Switzerland: RotoVision SA.

Durrant, P. and Doherty, A. (2010). "Are high frequency collocations psychologically real? Investigating the thesis of collocational priming." *Corpus Linguistics and Linguistic Theory* 6 (2). 125-155.

Elliot, A. J. (2015). "Color and psychological functioning: a review of theoretical and empirical work." *Frontiers in Psychology* 6, April 2015. 1-8.

Ellis, R. (2008). *Second Language Acquisition*. Oxford: Oxford University Press. 2nd edition.

Ender, A. (2016). "Implicit and Explicit Cognitive Processes in Incidental Vocabulary Acquisition." *Applied Linguistics* 37 (4), 2016: 536-560. doi:10.1093/applin/amu051

Ethnologue: Finnish language. Ethnologue.com. (10th March 2019.)

Filatova, O. (2016). "More Than a Word Cloud." *21st TESOL Journal.* 438-448. doi: 10.1002/tesj.251.

Flowerdew, L. (2015). Corpus-driven learning and language learning theories. In A. Lenko & A. Boulton (Eds.). *Multiple affordances of language corpora for data-driven learning*. Amsterdam; Philadelphia, PA: John Benjamins. 15-36.

Gries, S. (2013). "50-something years of work on collocations: What is or should be next." *International Journal of Corpus Linguistics* 18 (1). 137-165. Doi: 10.1075/ijcl.18.1.09gri

Halliday, M.A.K. and Mathiessen, C. (2014). *Halliday's Introduction to Functional Grammar*. London and New York: Routledge. 4th edition.

Held, B. (2018). "Positive Psychology's A Priori Problem." *Journal of Humanistic Psychology* 58 (3). 313-342.

Hsu, A., Chater, N. and Vitanyi, P. (2011). "The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis." *Cognition* 120. 35-55. DOI: 10.1111/tops.12005

Hsu, A., Chater, N. and Vitanyi, P. (2013). "Language Learning From Positive Evidence, Reconsidered: A Simplicity-Based Approach." *Topics in Cognitive Science* 5 (2013) 35–55.

Hunt, A. and Beglar, D. (2005). "A framework for developing EFL reading vocabulary". *Reading in a Foreign Language* 17 (1). 23-60.

Hyland, K. and Tse, P. (2007). "Is there an "Academic Vocabulary?" *TESOL QUARTERLY* 41 (2). 235-253.

Hyland, K. (2009). *Academic Discourse: English in a Global Context.* London and New York: Continuum International Publishing Group.

Hyland, K. (2012). "Bundles in Academic Discourse." *Annual Review of Applied Linguistics* 32. 150-169. doi: 10.1017/S0267190512000037.

Ivanic, R. (2004) "Discourses of Writing and Learning to Write." *Language and Education*. 2004, 18 (3). 220-245.

Kennedy, C. and Miceli, T. (2017). "Cultivating effective corpus use by language learners." *Computer Assisted Language Learning* (30) 1-2, 91-114. *DOI: 10.1080/09588221.2016.1264427*

Kurt, S. and Kingsley Osueke, K. (2014). "The Effects of Color on the Moods of College Students." *SAGE Open*. January-March 2014. 1-12.

#Lancsbox: Lancaster University Corpus Toolkit. http://corpora.lancs.ac.uk/lancsbox/ Accessed March 2019.

Latin is simple: texo, texis, texere C, texui, textum. https://www.latin-is-simple.com/en/vocabulary/verb/6698/. Accessed February 2019.

Laufer, B. and Aviad-Levitzky, T. (2017). "What Type of Vocabulary Knowledge Predicts Reading Comprehension: Word Meaning Recall or Word Meaning Recognition?" *Modern Language Journal* 101, (4). 729-741.

Lee, D. and Swales, J. (2006). "A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora". *English for Specific Purposes* 25 (2006). 56-75.

Lew, R. (2009). The Web as Corpus versus Traditional Corpora. In Baker, P. (ed.) *Contemporary Corpus Lingusitics*. London: Bloomsbury. 289-300.

Macis, M. and Schmitt, N. (2017). "Not just 'small potatoes': Knowledge of the idiomatic meanings of collocations." *Language Teaching Research* 21 (3). 321-340.

Mauranen, A. (2010). "Features of English as a lingua franca in academia." *Helsinki English Studies 2010* (6). 6-28.

Meara, P. and Miralpeix, I. (2016). *Tools for researching vocabulary.* Multilingual Matters.

Mesthrie, R. (2009). *Introducing Sociolinguistics.* Edinburgh: Edinburgh University Press. 2nd Ed.

Mokyr, J. (2005). *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton, NJ: Princeton University Press.

Nagy, W. and Townsend, D. (2012). "Words as Tools: Learning Academic Vocabulary as Language Acquisition." *Reading Research Quarterly* 47 (1). 91-108.

Nation, I.S.P. (2001). *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

Nation, I.S.P. and Webb, S. (2011). *Researching and Analysing Vocabulary.* Boston, MA: Heinle, Cengage Learning.

Norem, J. and Illingworth, K.S. (2003). “Mood and performance among defensive pessimists and strategic optimists.” *Journal of Research in Personality* 38 (2004). 351-366.

Norem, J. (2007). “Defensive Pessimism, Anxiety, and the Complexity of Evaluating Self-Regulation”. *Social and Personality Psychological Compass* 2(1) (2008). 121-134.

Nurmukhamedov, U. (2015). *An Evaluation of Collocation Tools for Second Language Writers.* Ph.D. Thesis, Northern Arizona University.

Online Oxford collocation dictionary: coffee. http://www.freecollocation.com/search?word=coffee.

Opetushallitus (2016). Opetussuunnitelman perusteet. https://www.oph.fi/ops2016/perusteet

Pajak, B., Fine, A., Kleinschmidt, D. and Jaeger, T.F. (2016). “Learning Additional Languages as Hierarchical Probabilistic Inference: Insights From First Language Processing.” *Language Learning* 66 (4). 900-944.

Pass, S. (2004). *Parallel Paths to Constructivism.* Greenwich, Connecticut: Information Age Publishing.

Pinker, S. (2007). *The Stuff of Thought*. New York: Penguin Group.

Pouget, A., Beck, J., Wei, J.M. and Latham, P. (2013). “Probabilistic brains: knowns and unknowns.” *Nature Neuroscience.* 2013, 16 (9). 1170-1178.

Pyrko, I., Dörfler, V. and Eden, C. (2017). “Thinking together: What makes Communities of practice work?” *Human relations* 2017, 70 (4). 389-409.

Rowley-Jolivet, E. (2017). “English as Lingua Franca in research articles: the SciElf corpus.” ASp [Online], 71 | 2017, Online since 01 March 2018, connection on 07 March 2018. URL: http://journals.openedition.org/asp/4987 DOI:10.4000/asp.4987

Römer, U. (2011). “Corpus Research Applications in Second Language Teaching.” *Annual Review of Applied Linguistics* 31. 205-225.

Sardinha, T.B. (2012). “Lexicogrammar.” In Chapelle, C. (Ed.) (2013), *The Encyclopedia of Applied Linguistics*. Hoboken, NJ. Blackwell Publishing Ltd. DOI: 10.1002/9781405198431.wbeal0698

Saville-Troike, M. and Barto, K. (2017). *Introducing Second Language Acquisition*. Cambridge: Cambridge University Press.

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. New York, NY: Palgrave Macmillan.

Scott, M., and Tribble, C. (2006). *Textual Patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Simpson-Vlach, R. and Ellis, N. (2010). "An Academic Formulas List: New Methods in Phraseology Research." *Applied Linguistics* 31 (4). 487-512.

Staples, S., Egbert, J., Biber, D. and Gray, B. (2016). "Development at the University Level: Phrasal and Clausal Complexity Across Level of Study, Discipline, and Genre." *Written Communication* 33 (2). 149-183.

Swain, M., Kinnear, P. and Steinman, L. (2010). *Sociocultural Theory in Second Language Education.* Bristol: Multilingual Matters.

Tenenbaum, J, Kemp, J., Griffiths, T. and Goodman, N. (2011). "How to Grow a Mind: Statistics, Structure, and Abstraction." *Science* 2011 (March 2011).

Thiessen, E., Girard, S. and Erickson, L. (2016). "Statistical learning and the critical period: how a continuous learning mechanism can give rise to discontinuous learning." *Wiley Interdisciplinary Reviews: Cognitive Science, Oxford.* 7 (4). 276-288.

Shawn, G., Milligan, I. and Weingart, S. (2013). "Voyant Tools" *The Historian's Macroscope - working title*. Under contract with Imperial College Press. Open Draft Version, Autumn 2013, http://themacroscope.org

Tono, Y. (2009). "Integrating Learner Corpus Analysis into a Probabilistic Model of Second Language Acquisition". In Baker, P. (Ed.) (2009). *Contemporary Corpus Linguistics.* London and New York: Continuum.

Tourmen, C. (2016). "With or Beyond Piaget? A Dialogue between New Probabilistic Models of Learning and the Theories of Jean Piaget". *Human Development* 2016, 59. 4-25. DOI 10/1159/000446670

Voyant Tools. (Website). https://voyant-tools.org/. Accessed multiple times 2018-2019.

Webb, S. (2008). "Receptive and productive vocabulary sizes of L2 learners." SSLA (30). 79-95. DOI: 10.10170/S0272263108080042

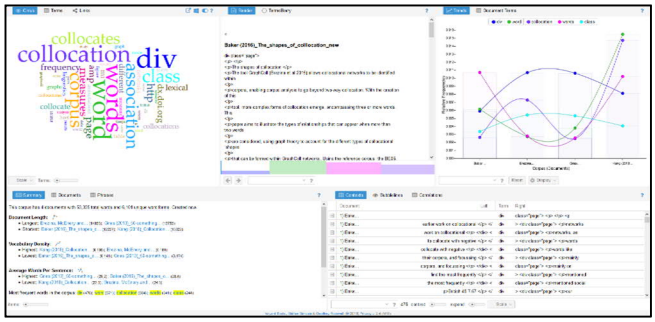
Weisser, M. (2016). *Practical corpus linguistics: an introduction to corpus-based language analysis*. Chichester, England: Wiley Blackwell. 1st Ed.

Witzel, N.O. and Forster, K.I. (2012). "How L2 Words Are Stored: The Episodic L2 Hypothesis." *Journal of Experimental Psychology: Learning, Memory and Cognition.* 2012, 38 (6). 1608-1621

Wu, Y., Provan, T., Wei, F., Liu, S. and Kwan-Liu, M. (2011). "Semantic-Preserving Word Clouds by Seam Carving." *Eurographics / IEEE Symposium on Visualization 2011*. DOI: 10.1111/j.1467-8659.2011.01923.x

Ädel, A. and Erman, B. (2012). "Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach." *English for Specific Purposes* 31. 81-92.

7. Appendix: The material Package



WORD CLOUDS AND BEYOND

Corpus Linguistic Material Package

LEARN HOW TO USE CORPUS TOOLS

Get more out of texts and develop your Academic English Writing

By Jere Hokkanen

Welcome!

In this material package, you will learn to utilize *corpus linguistics* – the study of text collections – as a resource for development as a reader and writer. The method for this are corpus programs: tools that do most of the manual work with texts for you. You do not need to be familiar with linguistic terminology, nor do you need to be a computer-savvy person: the course starts by familiarization to the programs, and linguistic jargon is avoided whenever possible. However, to be able to understand the process, you will learn some of the key terminology – you will need it to work with the programs efficiently.

During the course of this self-study package, you will learn to:

- Use two (!) different corpus programs
- Create numerous, different visual representations of any text you encounter
- Utilize these representations to analyze texts and gain more out of them
- **Create your own personalized corpus: a collection of texts from your own field of studies**
- Use your personalized corpus to understand the concepts in them better
- Recognize the norms of writing in your own field
- Utilize your personal corpus in academic writing
- Reduce your reliance on teachers and proofreaders
- Gain technical skill that you can utilize from now on: in studies, research, and even working life

Contents of the package:

You do not need to be familiar with corpus linguistics before using the materials presented here. However, there will be some unfamiliar terms that you will come by – they will be explained when they occur.

However, here is a handy reference sheet of some of the terms in case you need to come back to them.

Term in the material package	Meaning
Corpus	A collection of texts gathered for a specific purpose
Frequency	How often a word occurs in a text, or a corpus
Collocation	A link between two words, e.g. they often appear together. Even if they are not *immediately* next to each other. For example, black and white often occur together
a Collocate	A partner word in collocation: "*strong* is *a collocate* for the word *coffee*"
Concordance, or table of concordances	A table that shows the immediate textual context of a word: all the cases where the word appears and some words before and after it
Node word	The central word in a concordance table, or a word to which collocates are sought for
Network of collocations	A figure containing many node words and their collocates that are linked by lines
Concept	In this context, a word that refers to an academic concept important in a field: *discourse* in linguistics, *evolution* in biology, *affordance* in communication etc.
Lexical bundle	A bundle of words that has function in structuring the text: such "in the case of"

In the next page, there is a table of reference for all the tasks presented here. The columns, from left to right indicate:

1. The number of the task in the materials
2. The name of the task – this also tells you what is done in it
3. The program used in the task and whether it is something that can be accessed with an internet browser, or something that you need to download.
4. Learning objectives: what you learn after completing the task
5. Estimated duration – how long completing the task takes **at least**. A range is presented here (30-45 minutes, for example)
6. Whether the task is something that needs to be done in one go (**Not** continuous), or whether it is something that you can get back to relatively easily (**Yes)**

#		**Task**	**Program Used (online/ download)**	**Learning objectives** (what you will learn)	**Duration (estimate)**	**Continuous**
1		Word cloud Introduction	*Voyant Tools (*Online)	How to use a word cloud as a pre-reading tool for vocabulary	20 minutes + reading: 60 min total	No
2		Word cloud for Writing	*Voyant Tools* (Online)	How to make your writing more diverse	60-90 minutes (you need a text of your own)	No. works best as a follow up to task 1
3		Previewing concepts	*Voyant Tools* (Online)	How to use word cloud as a preview strategy to pay attention to how an author defines terms	90-120 minutes (work with Voyant, reading, writing)	Yes
4		**Creating a mini-corpus**	*Voyant Tools* (Online)	Tutorial on making a corpus with *Voyant Tools*. **Getting a mini-corpus to work with.**	30-90 minutes (depending on whether the user has the articles)	No, but mini-corpus is saved for further use
5		Concepts in a corpus	*Voyant Tools* (Online)	Use the options Voyant tools has for concept analysis. Learning to use new Voyant options.	90-120 minutes	Yes, but preferably not
6		Connections in a text	*Voyant Tools* (Online)	Using Voyant Tools to study collocations	90-120 minutes at least.	Yes, but preferably not

				from the mini-corpus. Learning to use new Voyant options	(90-120 minutes)	
7		**Cleaning up documents for a reference corpus**	Pdf editor, word or equivalent, Voyant Tools	Learn how to build a professional corpus of academic documents.	Roughly 10 minutes per text: 200 minutes total for a corpus of 20 texts	Yes, completely
8		Linguistic reference	*Voyant Tools* (Online)	Using the larger corpus as a linguistic reference	Practice 30 minutes	Yes, ongoing
9		Using Voyant Tools to study discourses	*Voyant Tools* (Online)	Use Voyant Tools as a tool in concept essay writing	Hours. Works best as a homework	Yes, essay
10		**AntConc: Concordance tables and lexical bundles**	AntConc (Download)	Learn a new corpus tool	60-90 minutes	No
11		AntConc as linguistic reference	*AntConc* (Download)	Learn how to use AntConc and your personal corpus to check language use	90 minutes	Yes
12		AntConc to improve your own texts	*AntConc* (Download)	Compare your own writing to professional writing and improve	90 minutes	Yes

Task 1. Word Cloud introduction

Time required: 20 minutes + 40 minutes reading = 60 minutes total. Has to be finished.

Explanation:

This task introduces you to an innovative tool utilizing the principles of corpus linguistics: *Voyant Tools*. The first task is simple. In it, you will learn to use a word cloud to preview an article you have to read. This is similar to a presentation: in presentations, you can see the topic(s) from the word cloud before it. This helps you to orient to the presentation. Here, the word cloud is used to preview the vocabulary of an article. If this task seems too easy, don't worry – more advanced tasks will follow. *Voyant Tools* can be used for much more!

Learning objectives:

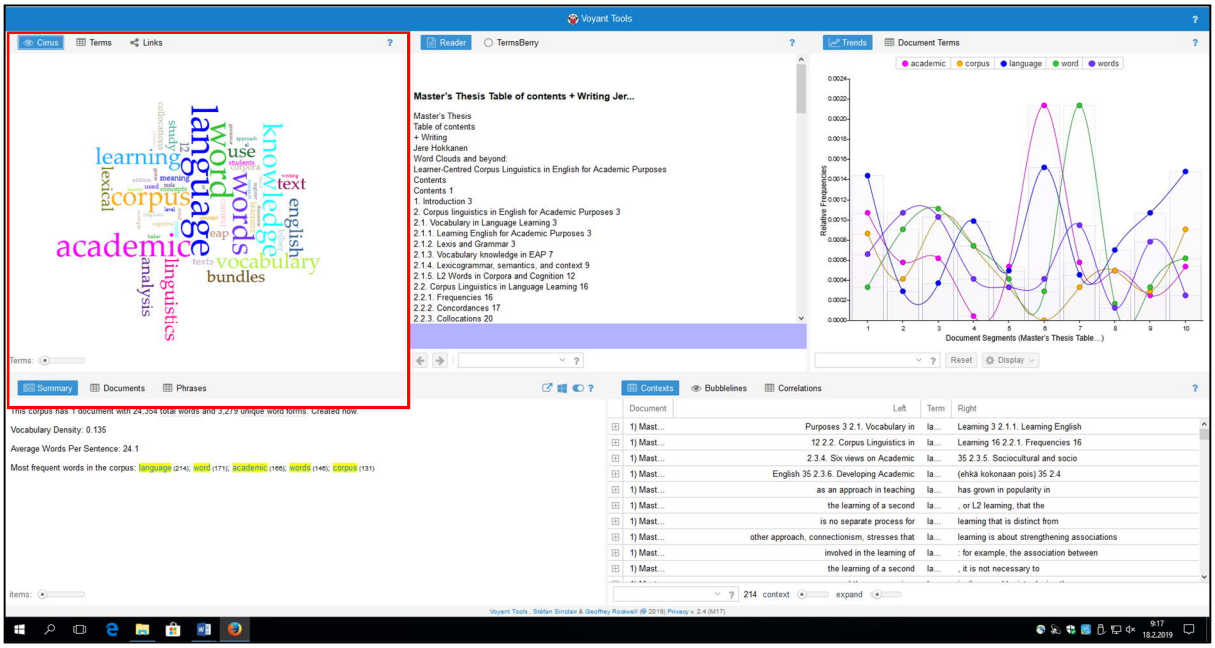
- Introduction to Voyant Tools
- Word cloud as a vocabulary preview method

Instructions:

Preparation:

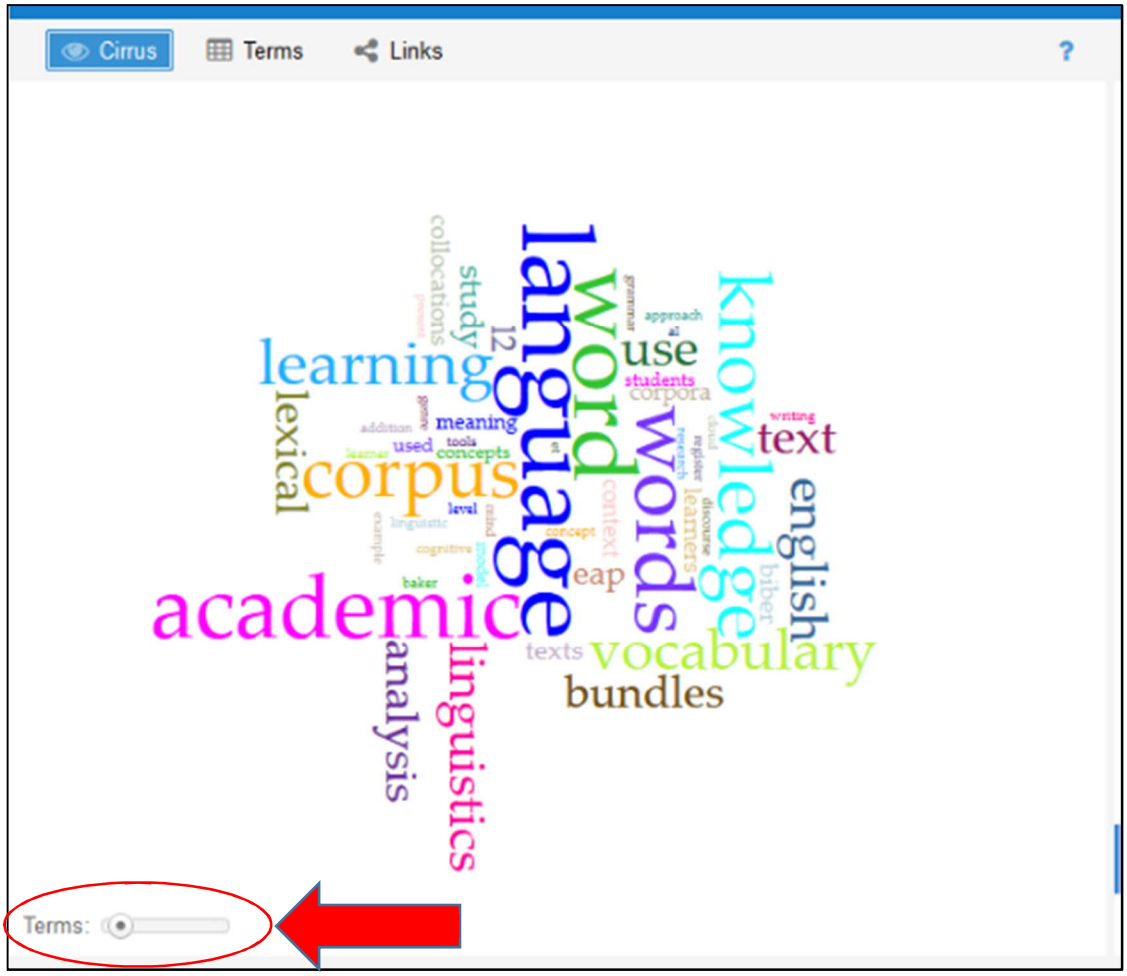
- Copy the text from an article you want to read with CTRL+A to select the whole text and CTRL+C to copy it.
    - OR download a text (.pdf, .docx)
- Go to https://voyant-tools.org/ and paste the text with CTRL+V to the box in the middle. Click on the box marked Reveal. The text will now be analyzed by *Voyant tools*!
    - OR click upload and find the article you want *Voyant Tools* to analyze

You will see the following *Voyant Tools* overview: the Word Cloud tools are marked inside the red rectangle:



Below, a closer look of the area marked with the red rectangle: *Cirrus*, *Terms*, *Links*: use this to select between **Word Cloud** view (*Cirrus*), *List of terms view*, and the *Links between words* view. Stay with *Cirrus* for now.

Terms slider: Use this to specify the number of Words in the Word Cloud. Try out different amounts and return close to minimum: to 50-75 words, whichever is the clearest.



The task:

1. Look closely at the Word Cloud. Are there any words that are unfamiliar to you, or that you are only somewhat familiar with? Before reading the article, go to https://www.merriam-webster.com/ and find out their definitions in English. Now read the article and pay close attention to how they are used in the article. Which definition is used?
- You may switch to **Terms** view above the word cloud, if it is clearer for you.

2. Congratulations! You have completed the first task.

Task 2: Word cloud for writing

Time required: 60-90 minutes. Has to be finished, best as a follow up to task 1.

Explanation:

In this task, you will use a word cloud to get information about how you write. This task is a follow up to task 1, provided you have a sample of your own writing to work with. In it, you will learn to notice the words you use most frequently in writing. Of course, the sample is only one article, but you can use this method to see afterwards, if your writing is too repetitive. After it, you can make some revisions and see the improvement.

3. Note: you need a file of a text you have written: an essay for example.

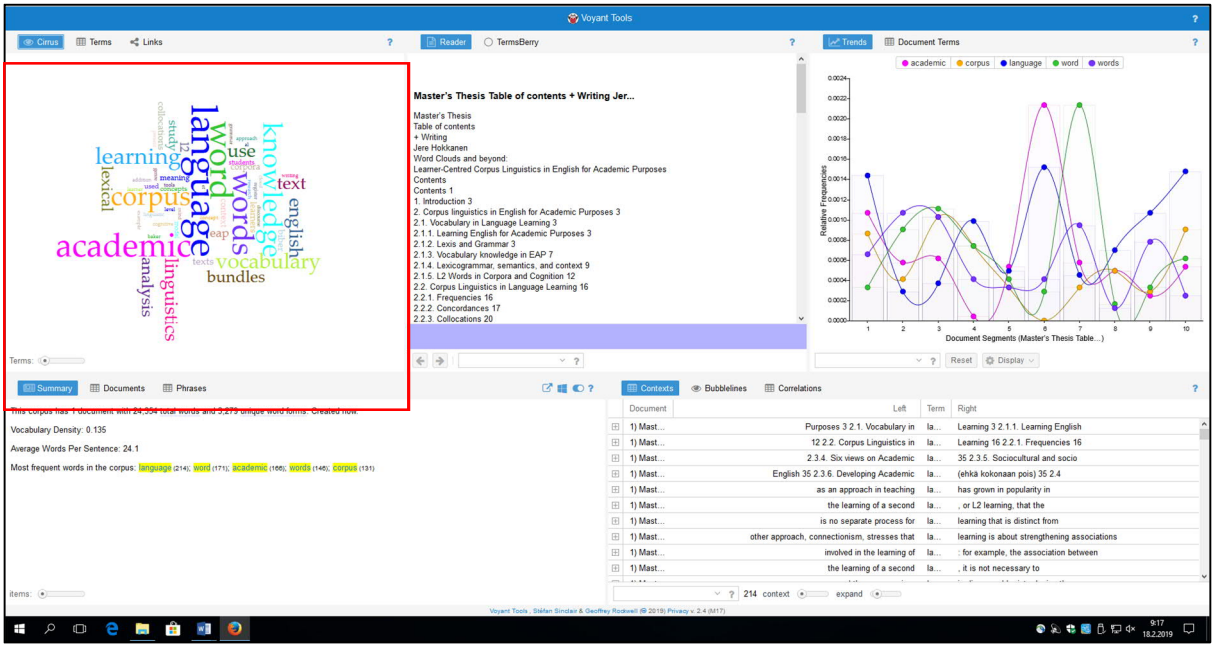
Learning objectives:

4. More practice with Voyant Tools
5. Notice the overuse of certain words in your, for example, if you always use “thus” to start a sentence that explains the previous one
6. Add variety your vocabulary

Instructions:

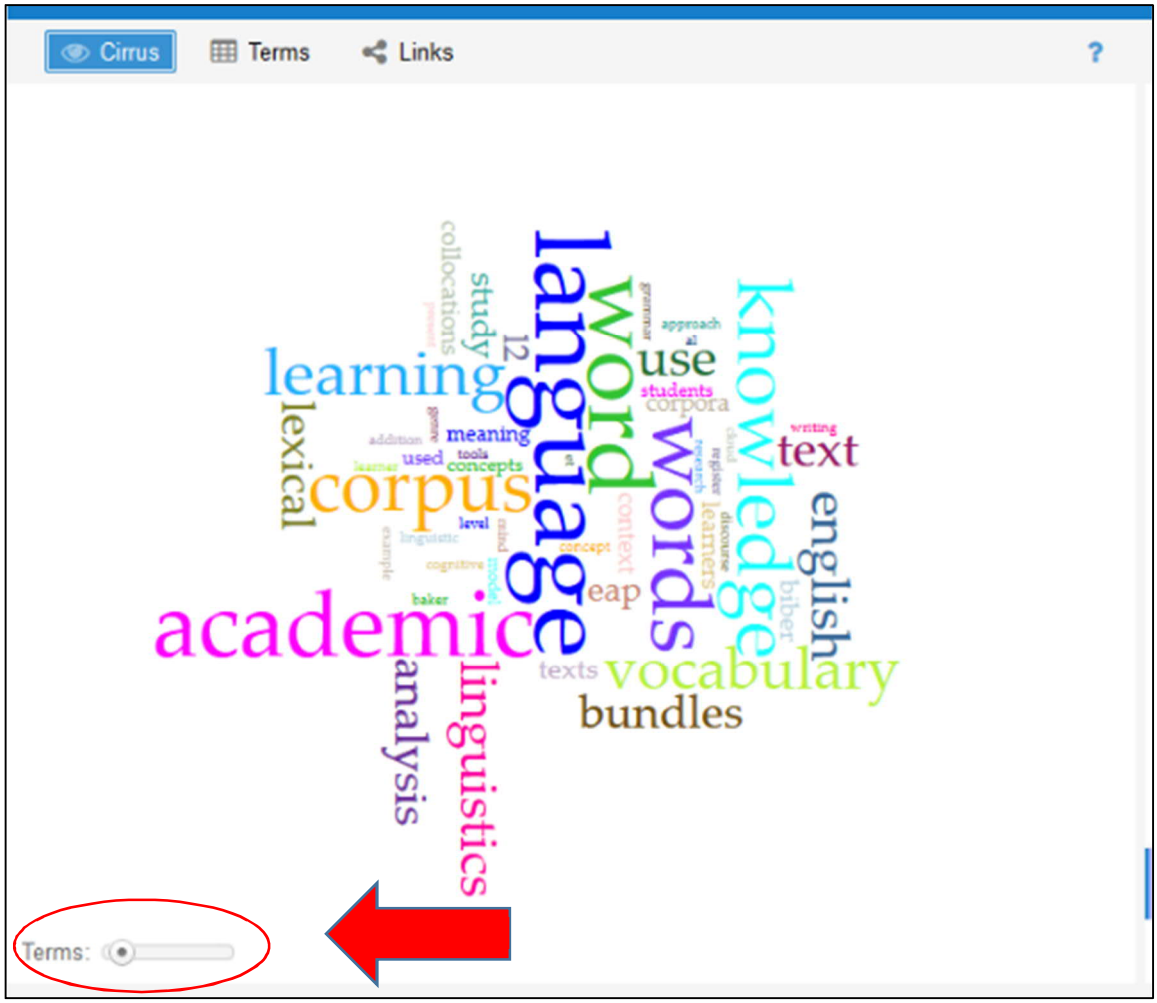
7. Go to the *Voyant Tools* page https://voyant-tools.org/,
8. Click on the box marked upload. Find an article you have written for analysis. The *Voyant Tools* analysis page will now open automatically!

You will see the following *Voyant Tools* overview: the Word Cloud tools are marked inside the red rectangle:



Cirrus, Terms, Links: use this to select between **Word Cloud** view (Cirrus), *List of terms view*, and the *Links between words* view. Stay with the Cirrus view for now.

Terms slider: Use this to specify the number of Words in the Word Cloud. Try out different amounts and return close to minimum: to 50-75 words, whichever is the clearest.



The task:

1. Look closely at the word cloud. What are the words that appear in the largest font? These words are the ones you use most frequently. Does the word cloud appear similarly balanced to the one in task 1?

- If the word cloud appears very different, you might need to improve your productive vocabulary!
- If you are not doing this task after number 1, you can just look whether the cloud looks balanced or not – if some words that you would not want to use too much appear in a larger font.

2. Pick at least 5 largest words from the cloud. Pick the ones that are not concepts or technical terminology. Go to https://www.thesaurus.com. Look at the alternatives presented in dark yellow.

- You can use **Terms** view to find out the largest – the five most frequent terms are also the largest in the cloud.
- Open each word in a new tab or window – you will need them later!

3. Open the article you have written on Word or the text program you use. Now, use CTRL+F to find out the words you have chosen in step 2, starting from the first. Find out all the occurrences in the text.
- Check the occurrences, going back to *Thesaurus* with each. Could any of these be changed with an alternative from *Thesaurus*?

4. Change some of the occurrences of each of the words with alternatives. Make sure you understand the alternative, so you do not misuse the word! Check https://www.merriam-webster.com/ if needed.

5. Save the modified article with a different name: for example, add _Voyant to the file name before the file type (.docx and so on)

6. Open *Voyant tools* again on a different window. Upload the modified file to it.

7. Look at the Word Cloud made by *Voyant tools* again.
    - If you are doing this after the first task, compare the cloud of your writing to the first one. Remember, the first article you had is not necessarily the best example of writing in its field.

8. Congratulations! You have completed the second task!

Task 3: Previewing concepts

Time required: 90-120 minutes. Can be returned to.

Explanation:

In this task, you will learn to use Voyant Tools identify the central concepts in a text before reading the article. This way, you will pay attention to the right things in a text. This is especially helpful for reading articles that are not necessarily ones you need to read carefully – you can read them for the most useful information.

Learning objectives:

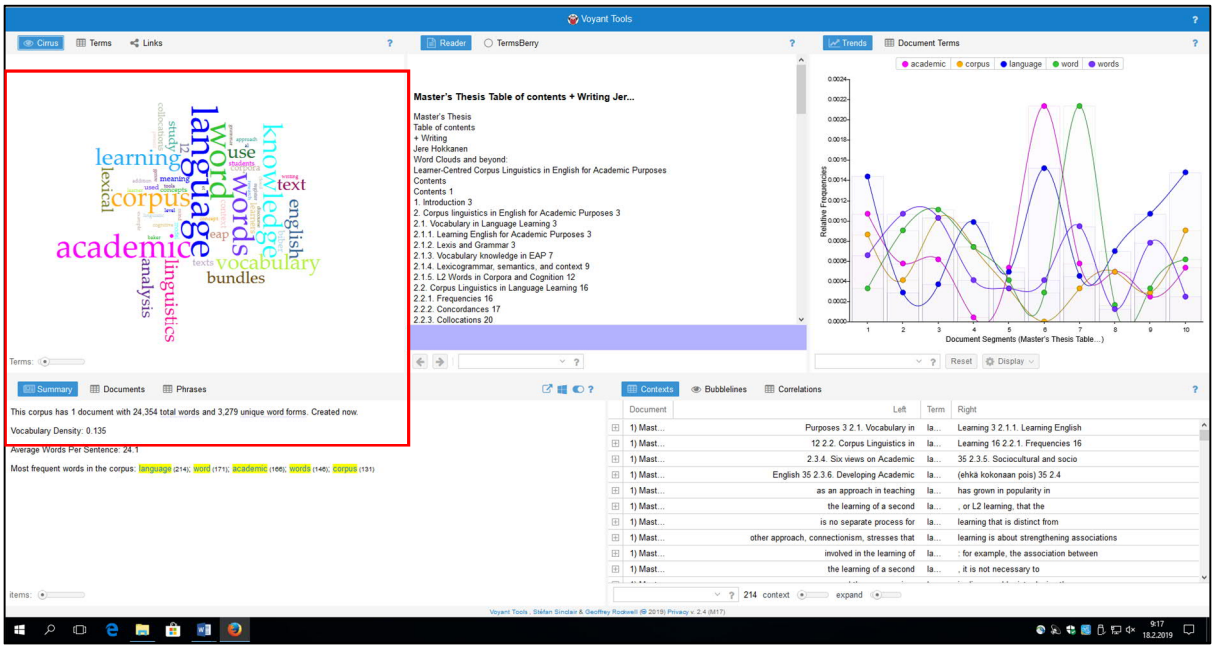
- Utilizing Voyant Tools as a pre-reading strategy
- Introduction to concept analysis using Voyant Tools

Instructions:

0. Pre-task: look at the *key words* of the article and read the abstract.

Pick a research article from your field of studies that you want to read. Download it on your computer if you have not done so already. Open *Voyant tools*. Upload the article.

Look at the word cloud view:



The Task:

1. Look closely at the word cloud. Adjust the amount of words if needed, or change to Terms –view. Pick the largest / most frequent *conceptual* terms from it, e.g. words that do not appear in typical everyday speech and denote a concept such as *discourse* or *nationalism*.
    - Pick 4-8 words.

2. Compare these words to the list of keywords for the article: any similarities or differences?

3. Spend about 30 seconds per word to think about their possible meaning(s).
    - If you are working with someone else, discuss the meaning of these words for about 1 minute per word.

4. Now, read the article and pay close attention to the words you have thought about or discussed.
    - How does the author use the word? What does it mean for him/her?

5. Write an essay where you explain the meaning of the words based on the article.
    - Write about one paragraph per concept.

6. **You are done! You have completed the third task, and the first three tasks. It will get a bit more complicated from now on.**

Task 4: Creating a mini-corpus with *Voyant Tools*

Time required: 30-90 minutes, you may need to look for the articles. Has to be finished.

Explanation:
In this task, you will learn to create and save a personal corpus of texts with Voyant Tools. This way, you can study the language from multiple documents and do many more things in later tasks.

Learning objectives:

- Tutorial on how to create a corpus of downloaded texts on your own with Voyant Tools!

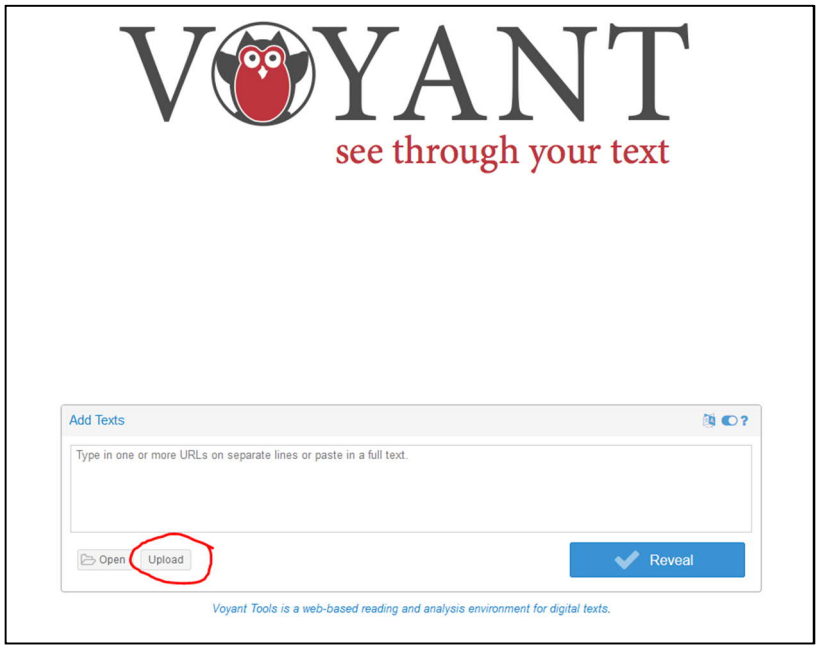
Instructions:
In the following example, I demonstrate how to use *Voyant Tools* to analyze three peer-reviewed articles that I have read and stored, dealing with the concept of *collocation*, the company a word keeps.

- Substitute a word you have articles for in your field of study here: a central concept in your studies!

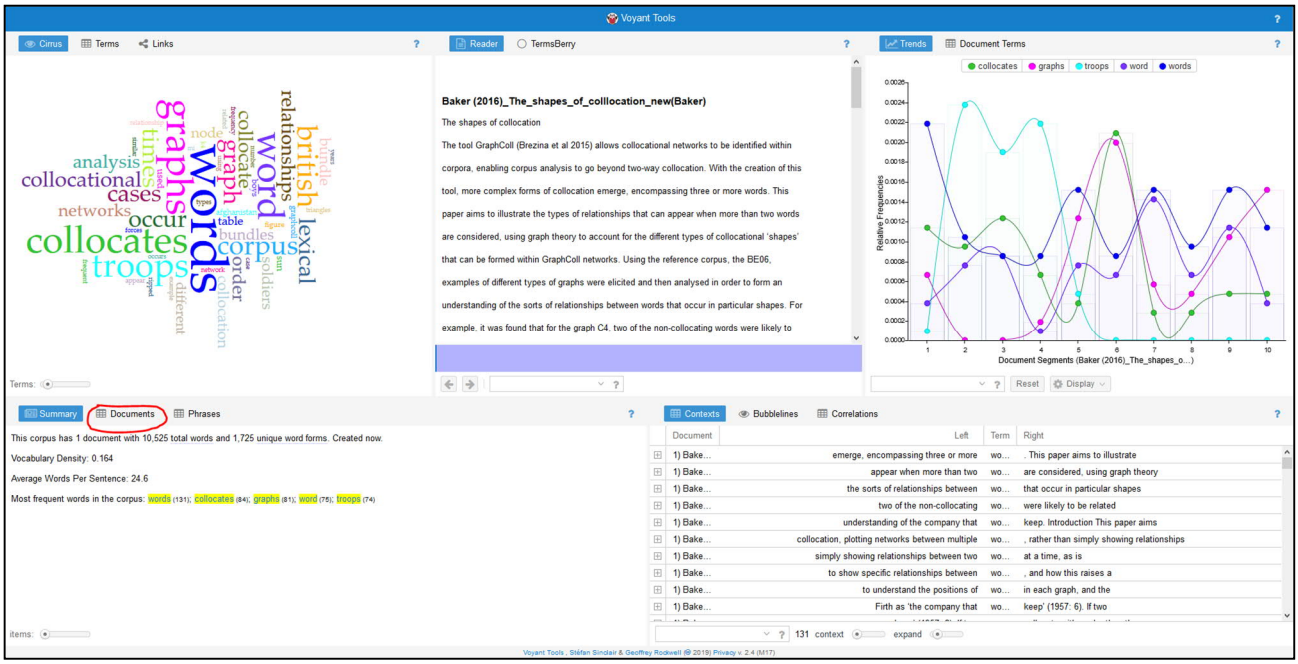
1. First, you will need to have the articles downloaded on your computer, or on a folder in a public computer such as a shared folder. Choose texts that you have read – for practice, this is simpler because you know what is in them. Pick four articles about the same same topic, and preferably of the same term/concept for now – for a central concept, you can use more, and for unimportant ones, less.

2. Now, open the *Voyant Tools* front page https://voyant-tools.org/

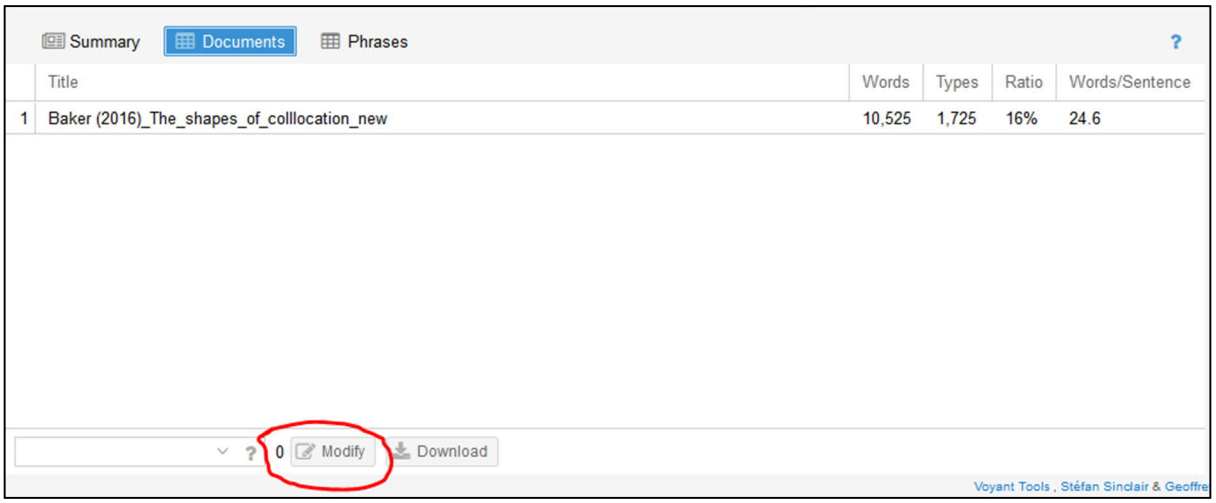
3. Select Upload:



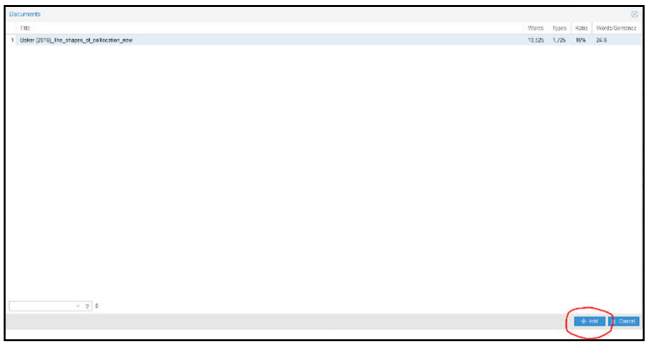
4. Now, choose the **first** file for the corpus. *Voyant Tools* now automatically opens the analysis page. Don't worry – you can add the other files later!



5. Open the Documents view marked with a red circle in the picture above. Then you will have the option of adding new files to your corpus. The view on the lower left changes to something like this:

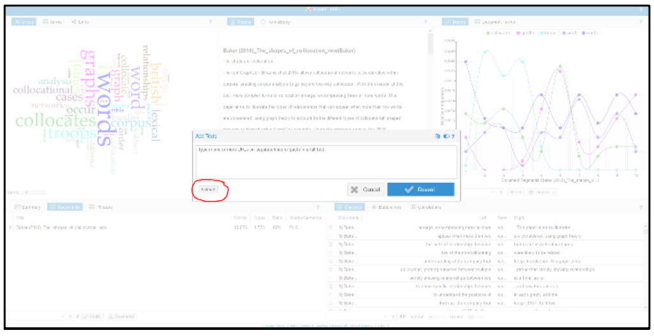


6. Click on modify. Another pop-up window opens:

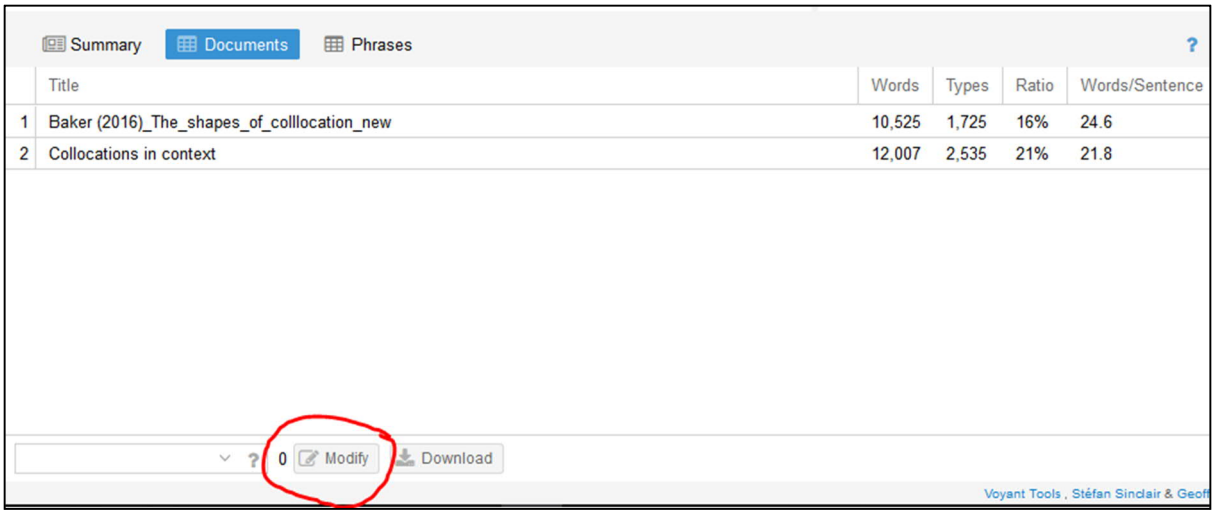


- For now, you only have the option of *adding* documents, because you only have one uploaded.

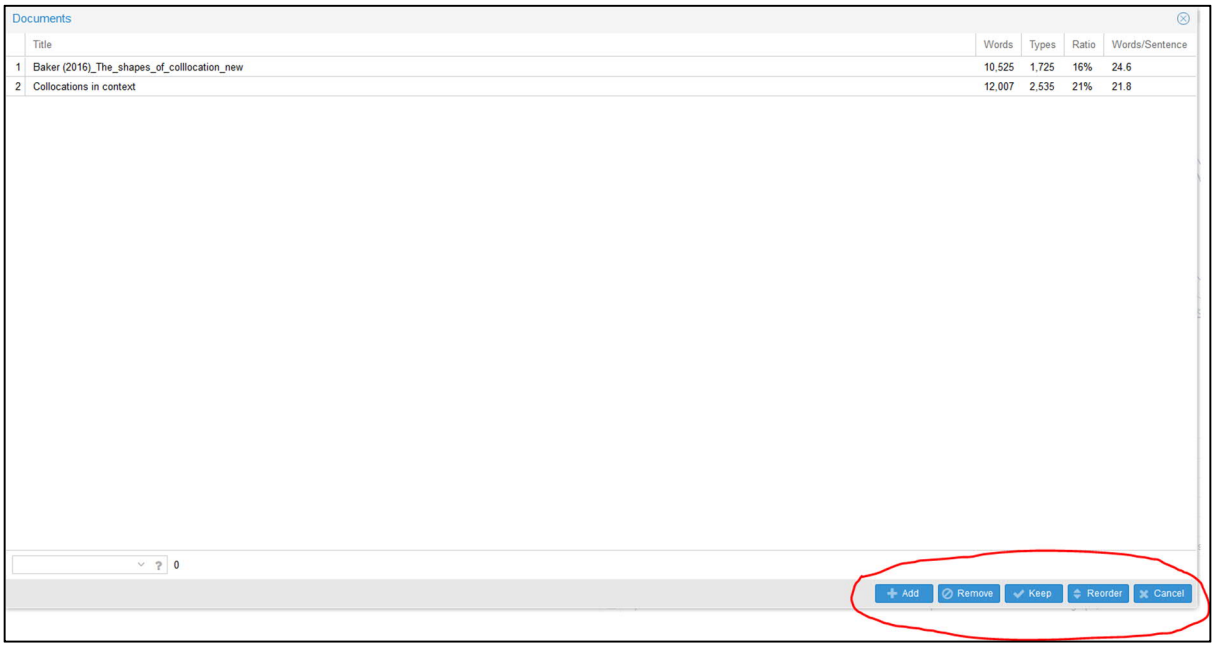
7. Now, click on Add. Then you will see another, familiar-looking pop-up window and the *Voyant Tools* on the background appears as faded:



8. Click on upload. Now, select the second article file you need for analysis. The view on the lower left changes: now, there is a second article and some basic information about it in addition to the first.
- Note that all the other parts of *Voyant Tools* have changed to fit the second article.

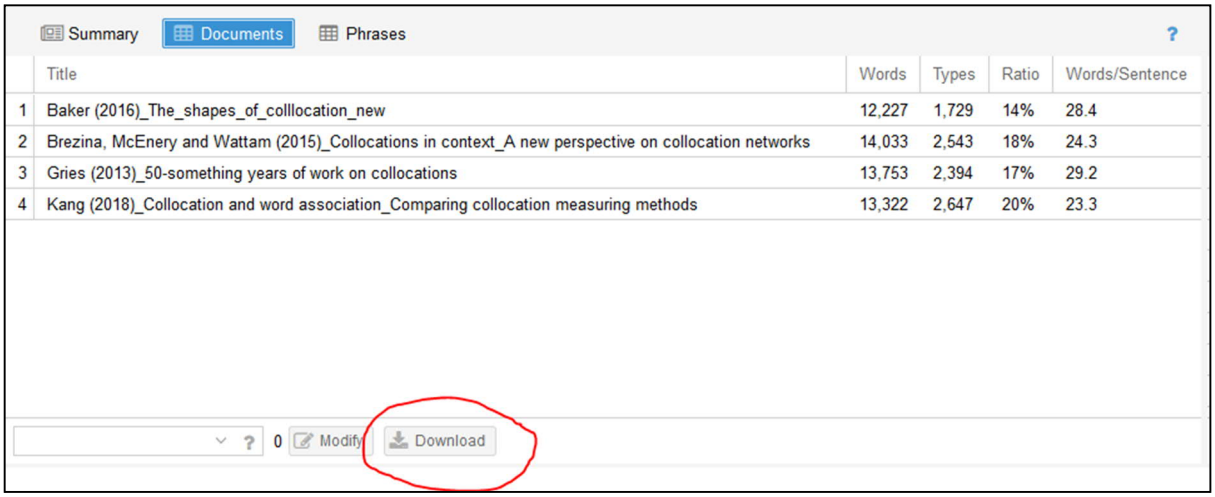


9. Congratulations! You now have a corpus of two texts. Now, add more texts to it. Click on modify again. Now, you also have the option of *reordering* and *removing* files from the corpus.

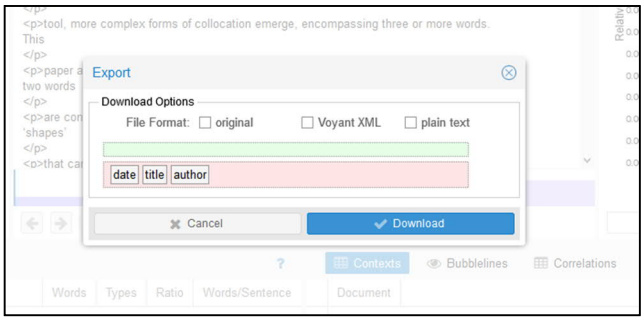


10. Repeat the process described in 6-8. Now, you have a corpus you can use to analyze multiple documents with *Voyant Tools*.

11. You need to save the corpus so you can use it for later reference. Click on the Download option:



12. Voyant Tools now shows different options for the download:



- You have the option of saving the files as plain text or Voyant XML, choose original (or do not choose any of the boxes, original is default). Then click download: Voyant Tools now provides a .zip file for saving.
- You do not need to unzip the file later for the individual – you can simply upload the zip directly to Voyant Tools later on.

13. Congratulations, you have successfully created your first corpus!

Task 5: Concepts in a corpus

Time required: 90-120 minutes. Best if finished in one session.

Explanation:

In this task, you will learn to use the mini-corpus you created and saved in the previous task to analyze multiple documents. The purpose here is to teach you how to find differences in definition of a central concept in your studies. As you may have noticed, it is common for different researchers to have their own definitions for concepts, which vary from each other.

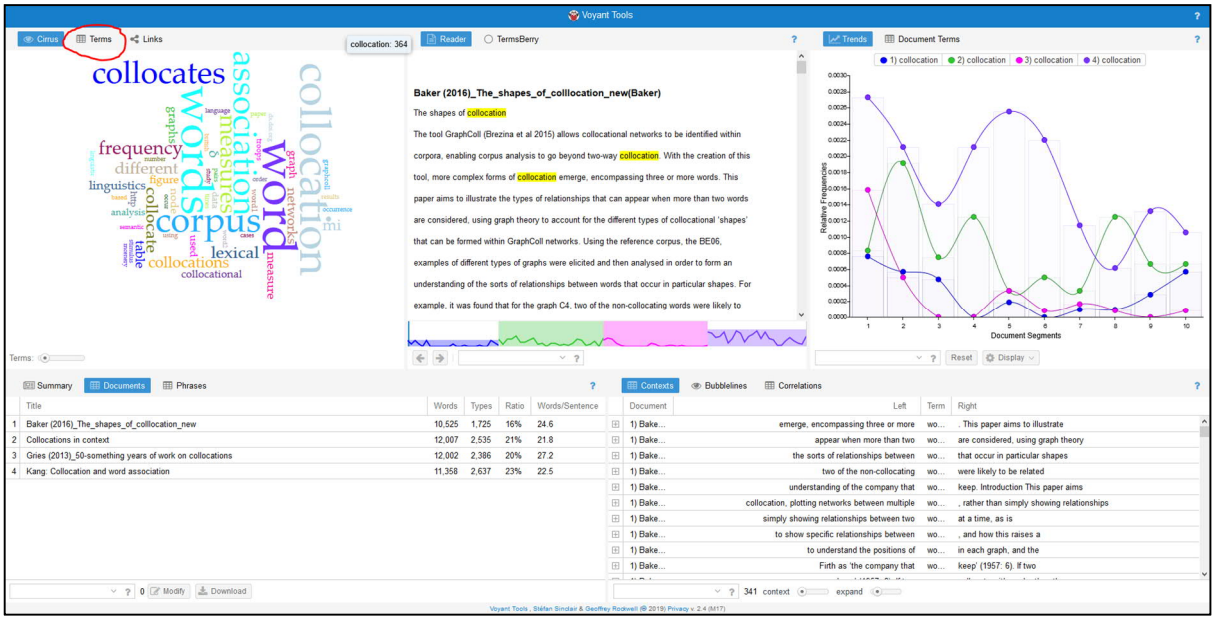
- Note that to develop expertise in the subjects that you are studying, you still need to carefully read the articles that are part of the courses in your studies, and the main articles and books you use for thesis work.

Learning objectives:

- Practice on working with the corpus you have created
- Learn about the different options Voyant Tools has for analyzing the language of multiple documents
- Reflect on the use of Voyant Tools for your own needs

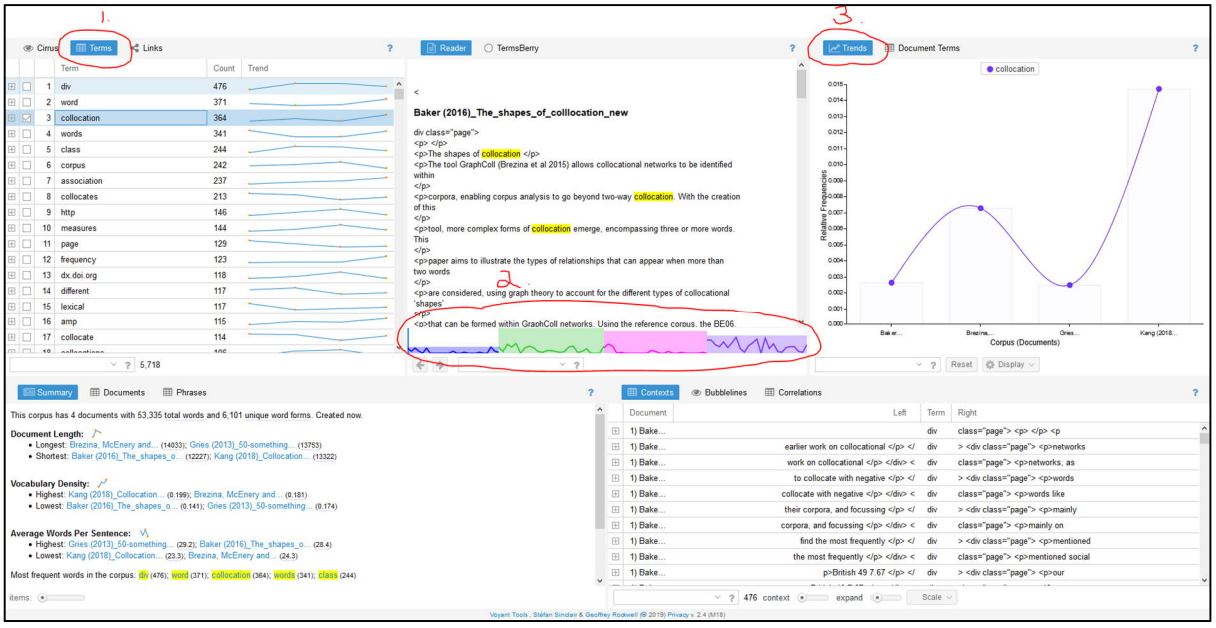
Instructions:

1. Upload the corpus you have created and saved in the Voyant Tools opening page. Click on the terms view, to switch to a list from the cloud.



2. The next step is to select a term for closer analysis. Because it is a central term in all the articles you have chosen for analysis, it should appear high on the list of most frequent words. In the example, I have chosen the word *collocation* for analysis.
   - Decide which four concepts you want to choose for analysis
   - Do not include basic, common words for the analysis here

- Start from one word
- In the example on the next page, I chose the word collocation for analysis



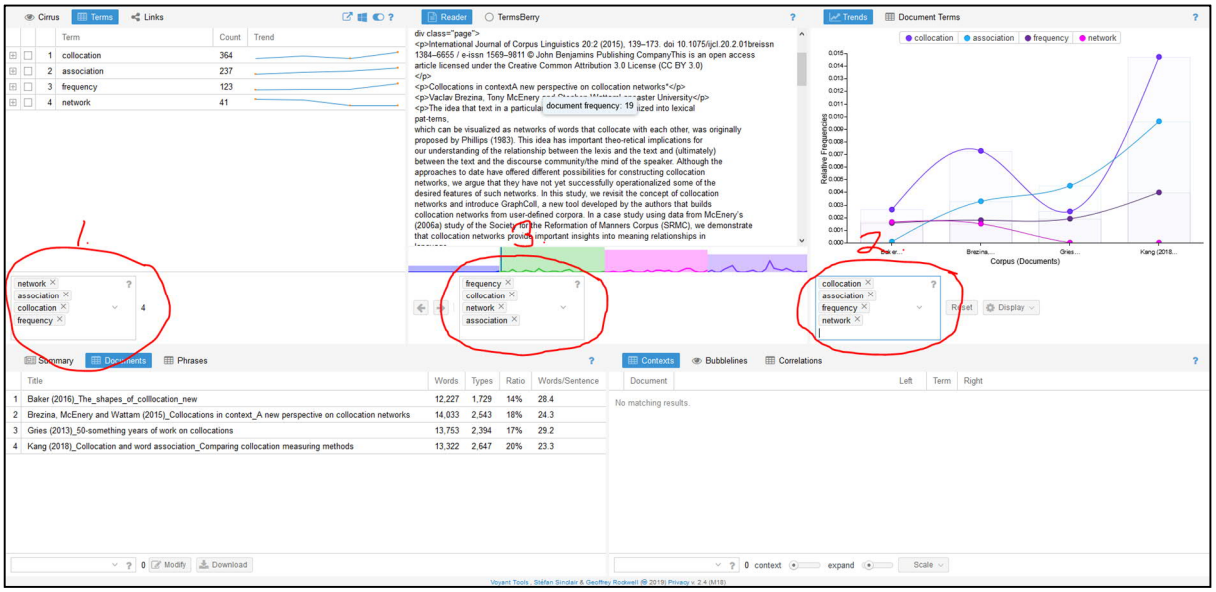
The picture above shows three graphs relating to the word *collocation*, marked with red numbers from 1-3.

1) The count and trend chart. This shows the number of times the word appears in the article, and the line shows where, and in which, of the articles the word appears most often. The articles are analyzed together as they were one text, in the order they were entered in *Voyant Tools*: you can see the order in the document view below.
2) The occurrence chart in the reader. Here, the articles are divided by color, and the chart shows where they particularly occur in. You can use the scroll function of the reader to find the terms: they appear highlighted in yellow.
3) The trends chart. You can easily see *where* and *how often* in the document the term you analyzing appears in – the y-axis graph "Relative frequencies" shows how often they appear. The X-axis marks the place in the corpora: each article is marked as a different segment.
   - Note that by default *Voyant Tools* picks the five most frequent words in the corpus for the trends chart. You can turn off the words that you <u>don't</u> want or need to analyze by clicking on them and their colored dot.

3. Explore different options and different combinations for a while: see how they affect the charts. Pick some basic words for comparison with the term/concept in addition to other terms.
   - *Voyant Tools* does not distinguish between actual words, numerals, abbreviations, and strings of letters like *http*. You will likely encounter them as well.

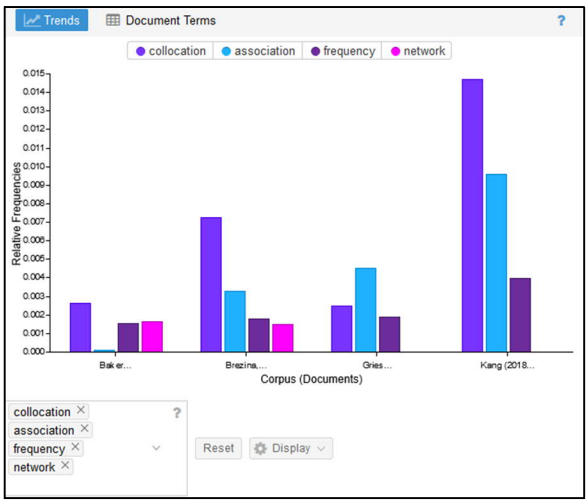
- *Voyant Tools* also lists the plural form of the term as a separate word. You should also mark the plural form for analysis!
- Remember, **you** need to do the actual analysis for your needs – the *Tools* is something you can use.
- You might encounter a lag with the *Tools* at this stage. If nothing else works, you may have to restart the analysis by closing *Voyant Tools* and opening the Corpus you have saved earlier in step 10.

4. Now, start making the observations more systematically. Enter the four terms you have chosen into the term selection boxes, in the terms and trends view.
   - In the screenshot below, I have entered four terms into terms in number 1, into trends number 2, and entered one term in the reader view in 3.
   - Note that the reader does not work with many terms simultaneously – you will need to do the analysis one word at the time with it. You can still write the terms there, it will be useful later.



5. Open a word file, or similar, so you can write down observations.
   - Start with the Terms chart. In Voyant Tools, The *Count* number reveals how many times the terms occurs in the corpus.
     - Note that the term has to be exact: you may need to include all variants of the term, such as the plural form.
   - Write your observations down, which term is the most common in the corpus. Include the numbers. Write as you would write an assignment – you will write all your observations into a brief essay.
   - Next, look at the line next to the *Count* number. By hovering your mouse cursor over segments of the line, you see the frequency of the term in each separate text in the corpus.
     - This is a handy tool for quick reference – but the actual *Trends* chart on the upper right side is much better, use it instead.
   - Now, move on to the Trends chart on the upper right, the tool to which you entered the terms in 2.

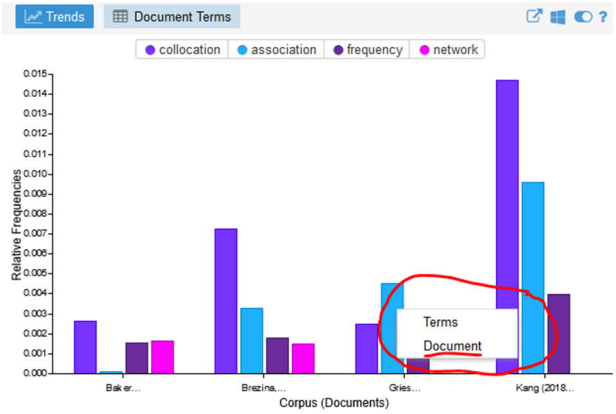
- The lines for each term show the relative frequency of them. The transparent bars indicate the amount.
- There are other visualization options that you can use, under the display option beneath the graphs. Lines and bars is the default.
    - Tip: Lines and bars is good choice for a large corpus that is, for example, arranged chronologically if you want to track how the use of a term increases or decreases.
- Switch to columns view. For comparing how frequently a term occurs in the different documents, it is a good choice, as shown below:



- By hovering your mouse cursor over a column, you gain the relative frequency information.
- By going back to the terms view in the upper left, you can turn on and off any of the terms from the viewer.
    - By selecting just one term, the Reader view and the graph in it updates to show it.

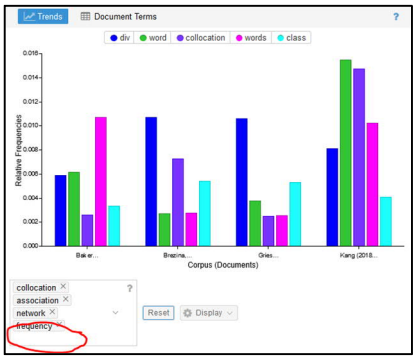
6. Study how often the words you have chosen occur in the documents. Write down your observations in the same word file you have used to mark down the frequencies.
    - Tip: note if some term does not occur in one or more of the documents.
        - In the case of documents that cover different matters, the chart may be very uneven.

7. Pay attention to the most frequent term in each document, the largest column in each segment.
   - Double-click on the column and switch to “Document” view:



   - This option is a bit buggy in Voyant Tools (at least by 3rd of April 2019). You will need to go to and turn off the other terms from the upper left to make it work. Otherwise, it shows false data.
   - Now, you can see *where* in the document term occurs in.
   - Study this, and include this information in your writing.
   - Go back to the general view.
     - Switching back to the general view is also a bit tricky: by clicking on reset, the trends view returns to the five most common terms in the corpus: but your selections are still visible.

   - To switch back, simply enter another word to the list (marked with a red circle) so that Voyant Tools updates the selections back, and then remove it.



8. Check your writing and make into a complete essay of about one page. Include a paragraph of reflection in the end:
   - What did you learn in the exercise?
   - How could you utilize Voyant Tools in the future?
   - **Congratulations! You have completed the fifth task. Next, you will learn about how words build connections in a text.**

9. Additional information: Trends chart
    - The horizontal X-axis shows the different articles/texts in the corpus.
    - The vertical Y-axis shows the *relative frequency* of the word in each text.
        - The number is calculated by dividing the word under analysis with the amount of the words in the article. For example: *association* / total number of words in Gries (2013) = 0,0045
        - This way, you can see how frequent and relevant the word under analysis is in each of the articles in your corpus. For example, the *trends chart* shows that the word "association" only once in the article by Baker (2016), close to zero in relative frequency.
        - In the case of a varied, large corpus, it is likely that many specialized terms will only appear in some of the articles – the trends chart helps you in finding out exactly which ones!
        - For example, the *Document* view on the word *association* shows that it occurs in the last segment. In the article under question, it occurs in the bibliography, in the name of a previous study.

Task 6: Connections in a text

Time required: 90-120 minutes. Best if finished in one session.

Explanation:

In this task, you will learn to utilize *Voyant Tools* to find out how the central words and terms are connected to each other in a text you want to analyze. This helps you to see relationships between concepts in the articles you are reading simultaneously – not just one article at the time.

Learning objectives:

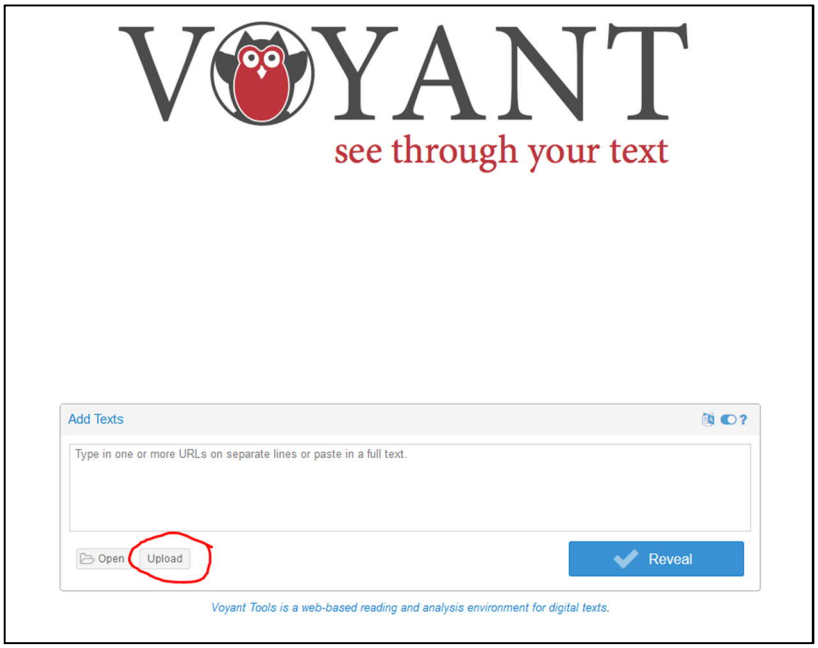
- Learn to use three new Voyant Tools options
- Learn a new Big Data method in studying how words are related to each other
- See how words are networked with each other in a text

Make sure you have access to the corpus you created and saved in task 4. If not, get this corpus, or complete **task 4** first.

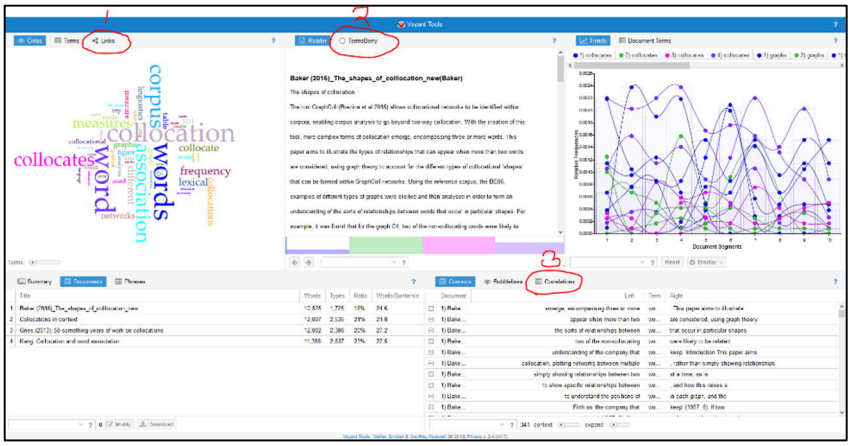
1. Open the *Voyant Tools* webpage at https://voyant-tools.org/

2. Upload your corpus to *Voyant Tools:*

3. Open a word document. Take notes on the links you see in the text throughout the process like you did in the previous task. This time, the instructions will not specify them more – you know how to do it.

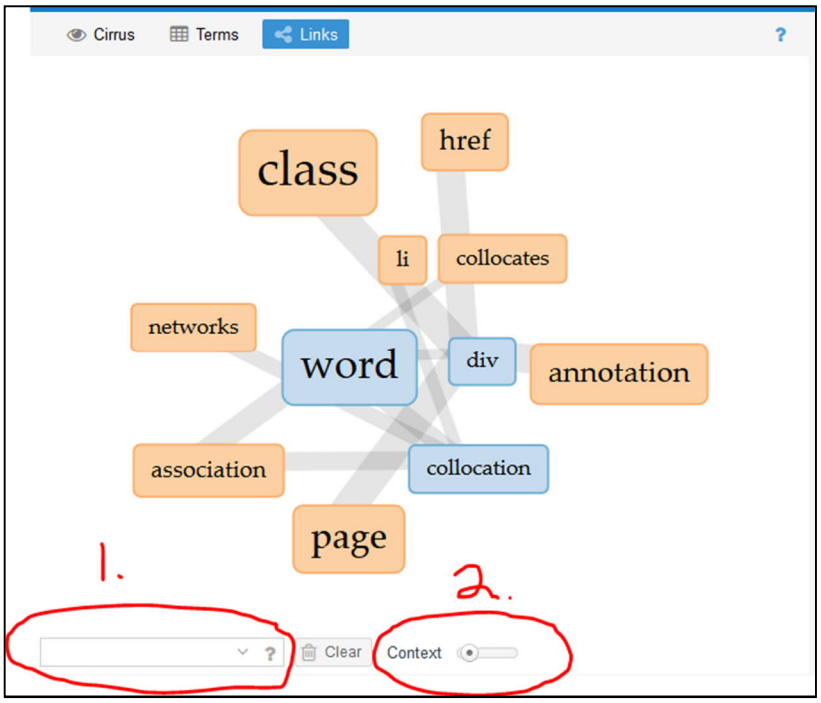


4. *Voyant Tools* now opens all the articles in your corpus for analysis. There are three options that you will be utilizing in this task, they are marked on the picture below:



1) First, there is the *Links* view next to the option of using *Cirrus* (Word Cloud), and *Terms* view. Links shows you which words link to the words you have chosen under analysis.
    - Remember, the default words under analysis in *Voyant Tools* are the five most frequent ones. You might want to change this later.
2) Second, there is *TermsBerry*: a handy tool that can track how many times a term occurs, in how many documents it occurs, and which terms it is linked to.
    - Because *TermsBerry* also keeps track of the number of documents a word occurs in, it works the best with a corpus and not so well with a single article.
    - TermsBerry will be covered last in this task, because the other two interact with each other.
3) Third, there the *Correlations* chart. It is used for more precise analysis of words linked together in the documents. The *Correlations* chart calculates the likelihood of words occurring together.

5. Start using the *Links* tool. For now, simply click on it. The view changes to this:



In the picture above, the words that have been chosen as the *node* words, the words that serve as starting points, are the ones colored light grey-blue, while the words linked to them are colored light orange. There are two options that you need to become familiar with.

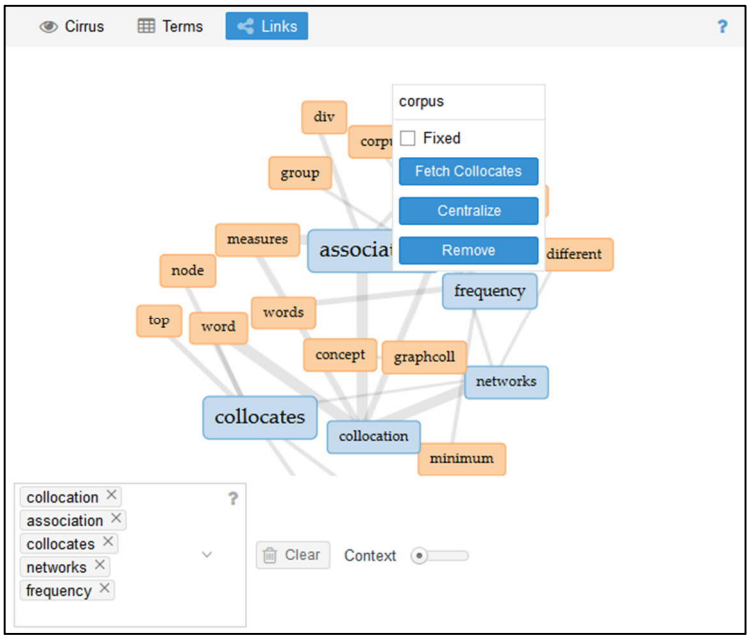
- **Tip:** sometimes resetting the *Voyant Tools* links view can be tricky. This is why it is advisable to have a saved corpus: you can reload Voyant.

1) The *search box*. By clicking on it, a list of the most common words in the documents opens up. You can also manually type a word there – and while typing, it suggests options that are in the document.
   - You can type/add multiple words here: the typed words will be the blue words, and the ones linked to them light orange
2) The *Context* slider. It allows you to expand the level of analysis: it automatically adds *node words* to the figure based on how common these words are in the text, starting from the most common ones.
   - The *context* slider **only adds.** This is why it is advisable to not experiment with the context slider too much.
3) In addition, there the *Clear* option marked with a garbage bin. It removes the words from the links view: this way you can reset the view.

6. Click on the *Clear* icon to remove the default words from the view.

- Enter, or select, some central concepts from the articles for closer analysis. In the example below, I have chosen the words *collocation*, *association*, *collocates*, *networks*, and *frequency* for closer analysis.
  - By hovering your mouse cursor over a word, it highlights the words and all the words that are connected to it in the figure.

- By right-clicking on a word in the network, you get several options for customizing the links view. In the example below, I have right-clicked on the word “corpus”

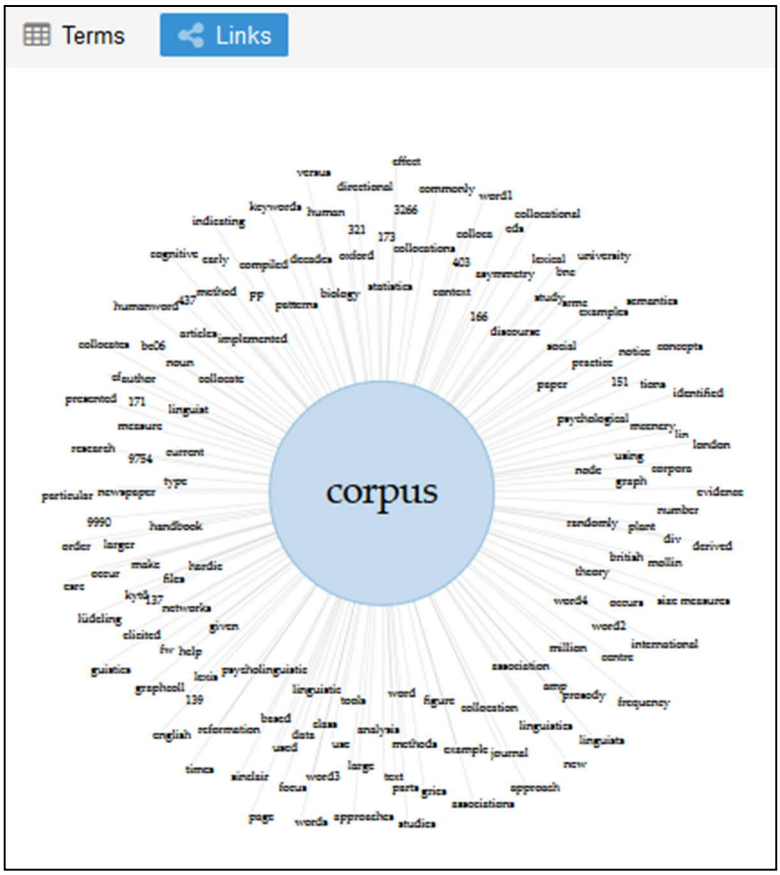


Four options appear:

1) Fixed or not fixed.
2) *Fetch collocates*: this option fetches the collocates for a word. With this option, you can pick a word that was <u>not</u> a word that you were originally searching collocates for into a node word. You can expand the level of analysis with this!
    - In the example above, the word *corpus* was a collocate of the word *frequency*. E.g. *corpus* was linked to frequency in the figure. By clicking on the “Fetch Collocates”, I can search for collocates to the word *Corpus* too. Now, additional words, the ones linked to the word *corpus* appear in the view in light orange, and *corpus* turns blue. Now I can find the collocates for the word *corpus*.
    - Try using this option now for one word!
3) *Centralize*. This option allows you to select a single word for closer analysis. Try this a bit later. Changing back to the earlier normal view is a bit tricky.
4) *Remove*. This option allows you to remove a word from the network. This is useful for removing non-words such as numerals, and words that are not useful for your analysis.
    - Try using this now to remove any non-words that may appear.

7. Closer analysis of a single word. Now, you will learn how to use *Voyant Tools* to see *all* the collocates the word has: all the words that are linked to it

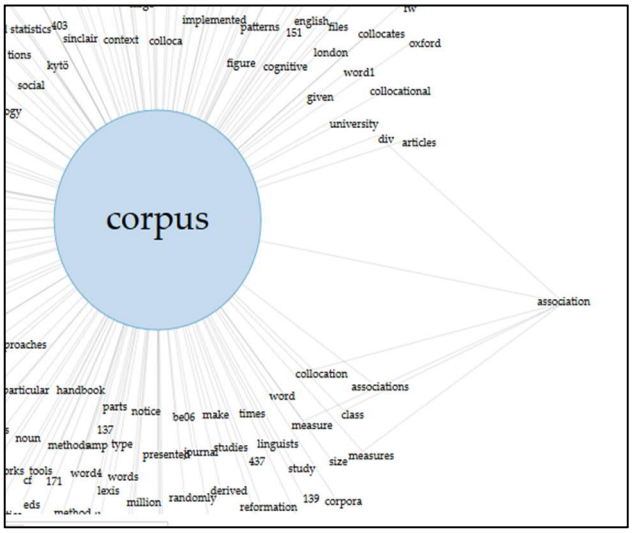
1) Select a word in the figure and right-click on it. A familiar view opens up: the one with four options. Now choose the option *Centralize*. The view changes:



- In the centralized view, the word under analysis is surrounded by all the words that are linked to it. You can use the mouse scroll to take a closer, or father look at the figure. In addition, by hovering the mouse cursor over by the words that surround the centralized word, you can see how many times the word links to the centralized word.
    - This amount indicates two things: first, how connected the word is to the one under analysis, and second, how common the word is in the corpus.
    - By double-clicking on a word, you can see **how** the word links to the central node word: either directly or via some other words. This view also allows you to see the words connected to the words you have clicked – once again expanding the analysis.

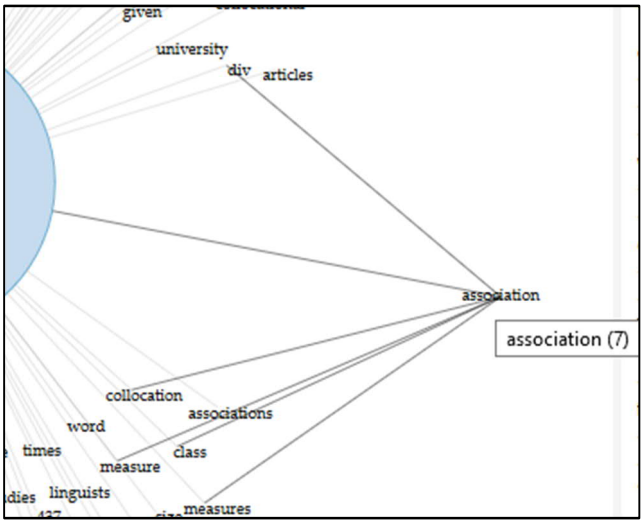
2) It is useful to know about the other options you have in this view: the sorting of words and a quick way to remove them.
    - You can sort the figure by moving the words around by clicking on them and holding the mouse button down.
    - You can remove the words by clicking on them, moving them outside the area they were originally in and releasing the mouse button. If you do not want to remove them, then do not release the mouse button when the text “Release to remove this item” appears – drag them back closer to the original word first.

3) Now, find a word among the surrounding words to see how it links to the centralized node word. Double-click it: in the example below, I have clicked on the word *association*:



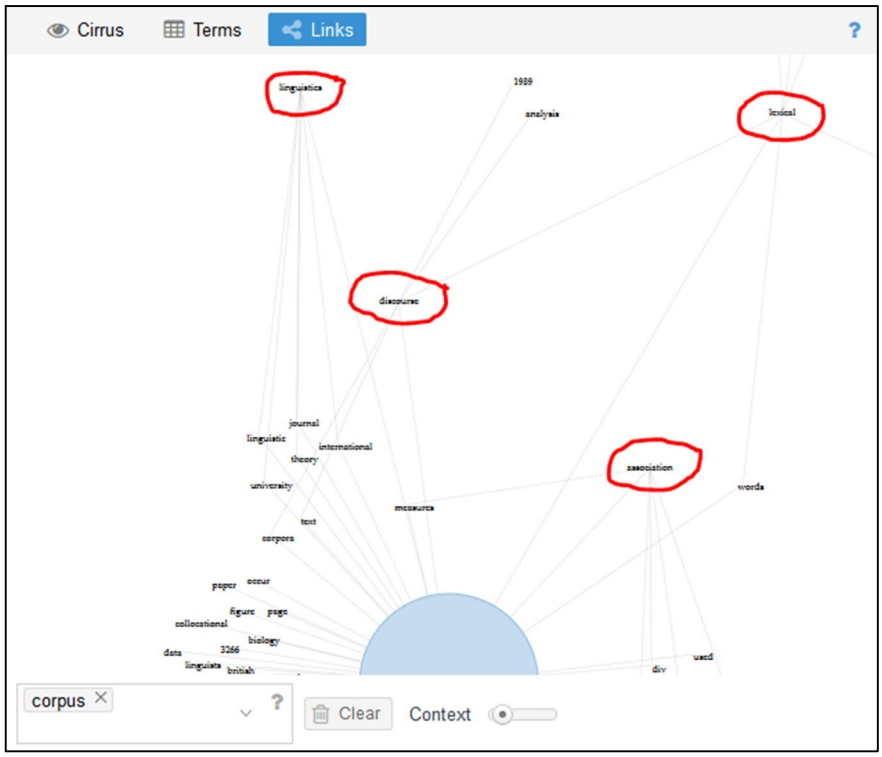
- Notice how *association* is set apart from the other words surrounding the *Corpus* circle: *association* is further removed from *Corpus*.

4) Hover the mouse cursor over the word you have chosen:



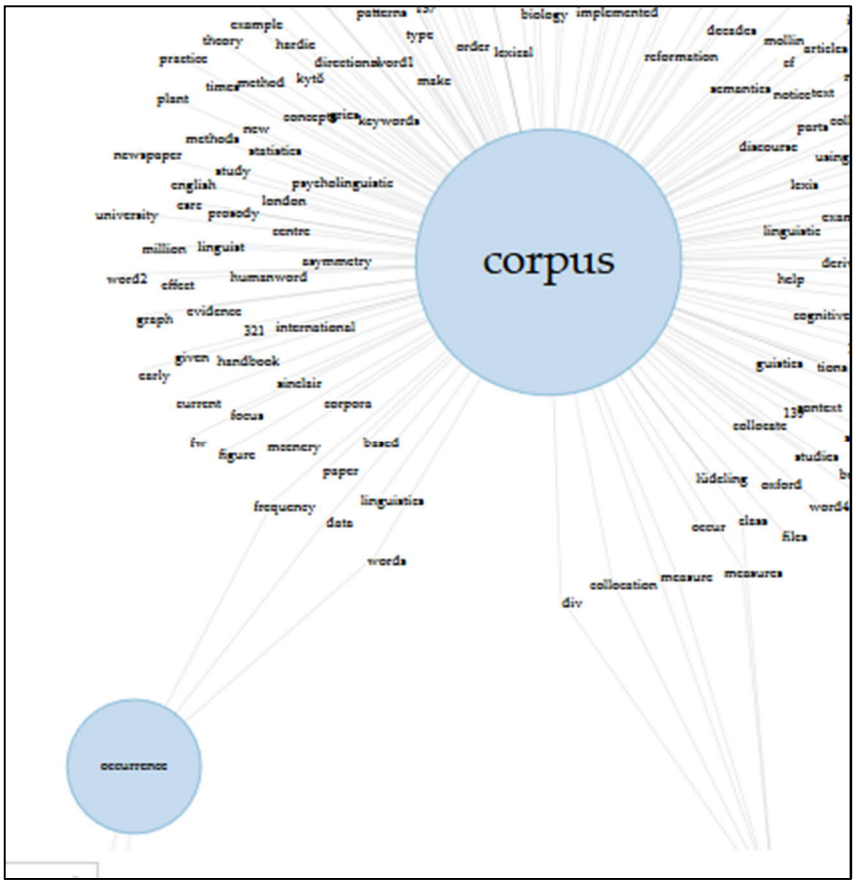
- You see additional information. You see where the word you have clicked links to and the number of linkages it has with the original word, in this case *Corpus*. This particular word links to words that also link to *Corpus*. It is possible for the word to be linked to other words that were not originally in the figure: then the figure expands to accommodate the new words.
    - You can also do the process explained at the end of step 4 for any of the words here: you can switch to the basic view with *Fetch Collocates*, choose another word to be *Centralized*, or *Remove* a word – even the centralized one (but it gets a bit weird if you do).

5) Do this for <u>four words</u> linked to the original word.
   - In the example below, I have chosen the words *association*, *linguistics*, *discourse*, and *lexical*.



   - As seen in the screenshot, some words link "outside" the original figure and expand the view.
     - In the case of a large corpus and/or a word with many linked words, the text in the figure gets small – you need the zoom feature!
   - A word of caution. The centralized view can get a bit tricky at times: sometimes double-clicking seems to hide the word if there are too many words open.

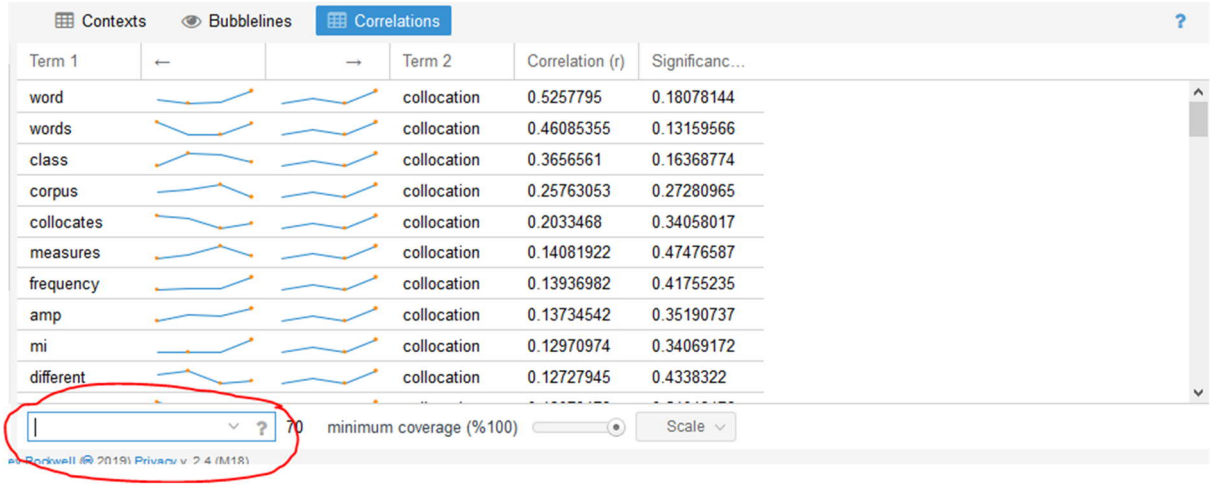
- You can also add another centralized word in this view by typing the word on the search box: in the example, I typed *occurrence*. This is a clearer way to analyze the mutual links between two words.



- For example, *Occurrence* and *Corpus* are connected via the word *Frequency*. Since frequency in Corpus Linguistics indicates how often the word occurs in a text or a corpus. In this case, the link makes sense.

8. Next, you will learn how to analyze collocations numerically with the *Correlations* chart.
    - This is a great tool if you prefer numerical data and statistically significant correlations: all the numbers are visible there and Voyant Tools provides you with information about whether a correlation is significant.

- Type the word you want to search correlations for.
    - The words that correlate highly with the word are termed collocates – this is the same thing that you did with the links view, only numerical.
    - Note that the word 'disappears' after you have written it to the search box marked with red – it becomes *Term 2* in the chart.
    - In the example below, I have typed the word collocation, which was central in the corpus I have:



- The default option shows you the words arranged so that the term 1 with what the correlation is the strongest appears up in the list.
    - The blue dotted lines show the relative frequency throughout the documents in the corpus for both terms.
    - The correlation can be negative: this means that the presence of one word means that the other word is not likely present.
- Repeat the process for all the words you analyzed in the earlier sections.
    - Remember to take notes on particularly significant correlations – by combining this information with the visual on earlier parts, you get a good idea on how some of the words are linked together in an article.

9. **You are done! You have seen how words form connections in a text, and learned to use new Voyant Tools options.**

Task 7: Cleaning up documents for a reference corpus

Time required: 200 minutes, 10 minutes per text. Can be done in small increments.

Explanation:

A mini-corpus is good for studying and comparing many features of the documents gathered there, as you have seen in the previous tasks. However, if you want to use a corpus as a reference tool for how the academic English language works in your discipline, you need a corpus that contains more documents from your field. In this task, you will start building such a corpus.

This task is especially suitable for you if you are doing a thesis work or soon planning to do so, as then you have academic articles that work as documents for the reference corpus.

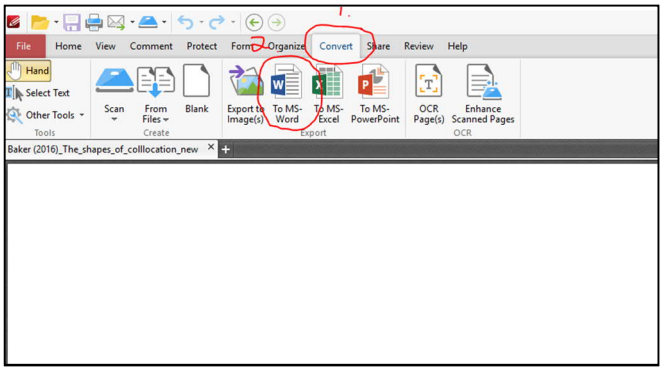
Learning objectives:

- Converting .pdf files into .docx, and .docx into .txt
- Learn how to clean up document for a proper corpus
- Practice on finding published articles in your field of studies
- Build your own reference corpus that can be used for Voyant Tools **and other corpus programs**

You can utilize the documents you chose to build a mini-corpus of for task four in this, but you need to clean them up first.

1. First, you should have or find **at least 20** peer-reviewed, published articles from your discipline / field of studies to work with.
    - The articles should be up-to-date, year 2000 and after is good enough here
    - Try to find documents from many authors, so that you get a good representation of different styles
    - Long articles and ones that do not contain lots of tables, figures, and visuals work the best here
    - If you cannot find the required amount just now, but can do so later, move on to step two. You can add more articles later.

2. When you have articles to work with, start by opening one with a pdf reader such as PDF-Xchange Editor.
    - Convert the pfd to a MS Word file so you can edit it more easily:



3. Save the converted file into a folder of its own.
    - You can label the folder something like *corpus_building*, so you know where to find all the converted files.

4. Open the converted file in Ms Word or equivalent. Now, you can start cleaning up the document for corpus use. The cleaning up means that you will remove all the data that is not part of the text. This should be done in a systematic way.
    - The document does not need to look nice and readable to a human eye after these steps: the point is making it readable to corpus programs.
    - Some documents require a lot more work, depending on the publisher. For linguistic reference, you might even want to skip these files.
    - **Note! If the amount of work seems excessive, do at least phase 1. The second most important is phase 4 – prioritize these if necessary.**

    1) First, remove the cover page, front matter, and the table of contents if the document contains such.
    2) Second, remove the citation guidelines (how to cite) from every page if necessary, and if there is the name of the publication on every page, remove it.
    3) Third, go through the document and delete all tables and figures.
        - Remember to remove the identifier of the table or figure too.
    4) Fourth, remove the entire bibliography and sources.
    5) Fifth, remove any appendices from the document.

5. Save the modified document.
    - Note that the document still might contain some information that is not purely textual, such as citations inside the text and numerals. This is fine, as they are not so frequent as to affect the corpus when more documents are added.

6. Make a second copy of the save with a different name with the *Save as* feature, put your name in it: *Familyname_Corpus.docx*
    - The point of making another copy is that the new file will be the corpus you will be building and you don't lose the separate, cleaned up versions of the documents, in case you want to use them to build a different corpus later.
    - You will be adding more documents to this document.
    - This is done so that the work

7. Now, you should add a second document to your cleaned up corpus. Repeat the process described in steps 2-5.

8. After saving the second cleaned up document, do not make a second copy of it like you did with the first. Keep just one copy. Then
   1) Select the whole text of the second cleaned up document with Ctrl+A and copy it with Ctrl+C.
   2) Open the (your) *Familyname_Corpus.docx* that you will be adding more documents to
   3) Scroll down to the end of the document and add the text of the second document to the end of the first.
   4) Save the now enlarged *Familyname_Corpus.docx*.

9. This is the process – now you will need to repeat it a minimum of 18 times.
   - You can get back to this process anytime if you do not have the time to do it now.
   - Remember to keep the cleaned up versions of each document. This is helpful if you need to make a second corpus, and also for tools such as *Voyant Tools* that you worked with previously that can distinguish between documents.
   - The *Familyname_Corpus.docx* is so for other programs like *AntConc* that you will work on in tasks 10-12, and GraphColl in tasks 13-15.

10. Now you should have 20 cleaned up documents, 20 cleaned up documents saved in your *corpus_building* folder, and a single file containing all of them.

11. **Congratulations, you have a professional corpus! Now you know how to make one, and add more files to a corpus!**

Task 8: Linguistic reference

- **Note! You cannot do this task properly if you do not have access to a linguistic reference corpus made in Task 7.**

Time required: 30 minutes. Should be finished in one session.

Explanation:

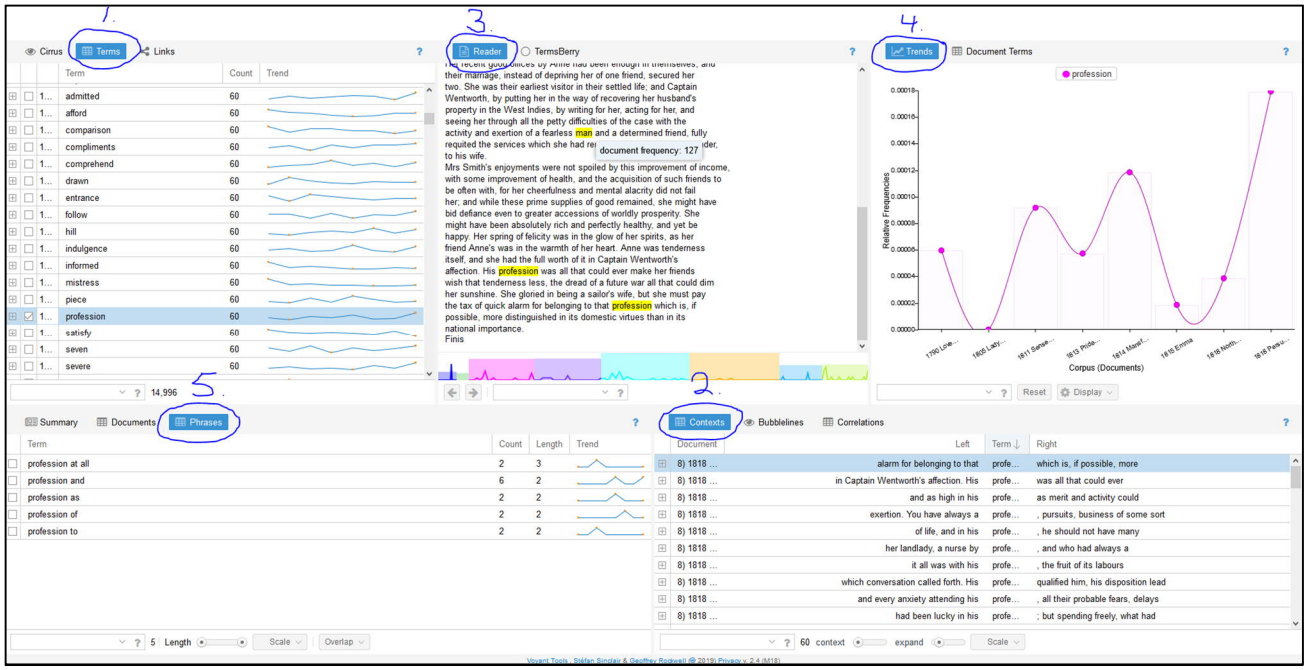
In this task you will learn how to use Voyant Tools as a linguistic reference tool – a tool that you can use to check on how professional writers use the words and in what contexts, what sort of prepositions (in, on, at…) go with certain words and so on.

Learning objectives:

- Learn how to use a corpus to check how language is used in actual contexts
- Gain familiarity with a central corpus linguistic analysis method: analyzing a *Key Word in Context, KWIC*.

1. Make sure you have access to the linguistic reference corpus you made in Task 7. Open the Voyant Tools front page https://voyant-tools.org/ and upload your corpus to Voyant Tools.
   - For Voyant Tools, you should do the process described in Task 4 and load all the documents separately to the corpus.
   - If, for any reason, you do not have access to this corpus you should open one of the default corpora on Voyant Tools. I demonstrate the use of these Voyant Tools options with the corpus of (Jane) Austen's novels, which is one of the two default options.

2. The Voyant Tools overview opens up. Use the tools marked with the *blue* circle and the number in the screenshot below for analysis
   - Psychological research (Elliot 2015) has revealed that the color blue helps with attention – as the tasks have gotten harder, you will need it!
   - I chose the word *profession* for analysis, as a *Key Word in Context*.
   - Screenshot in the next page!



1) You need to use the terms selection to find the term you want to analyze from the corpus (you can use the search feature if you know which one you are using). Mark the box next to the term you want to analyze, so that the trends figure in 4) and the phrases list in 5) update accordingly.

2) The *Contexts* option is the one you will be mostly using here. You need to write the term you want to analyze to the search function below. I have done it already so that the contexts table has *Profession* as a key term.
    - The scrollable table shows all the cases of the term occurring in the corpus, or a single document. *Scale* feature under the key words allows you to specify which documents you are looking the term from.
    - The table, called a *table of concordances* in corpus linguistics, shows some of the words around the KWIC.
    - This allows you to study where the word typically occurs in a sentence.
    - You can also see the prepositions for physical objects in the table.

3) The *Reader* option is handy with the *Contexts* option: whenever you click on one of the lines in the table, the reader finds the full context where the term occurs in.
    - Due to the selection you made in the terms, the term you are looking for is highlighted yellow. This also helps!

4) Trends chart is not part of this task so much, but again, you can quickly see in which documents the term you have chosen occurs in.

5) *Phrases* option shows you all the repeated phrases and multi-word units the term occurs in.
    - This is also an exceptionally good option for finding prepositions.

3. Practice using Voyant Tools as a linguistic reference with six terms. Pick different types of words (verbs, nouns, etc.), and both words that are common in general language use, and ones that are specialized.

4. Write brief notes on the usage of the words for yourself.

5. Start thinking about ideas on how to utilize the tools presented here and before for your own studies or as a research tool, if you study in the Humanities or Social Science.

6. **You are done! Only one task with Voyant Tools after this – then you will learn how to use a dedicated concordance program, AntConc.**

Task 9: Using Voyant Tools to study discourses

- **Note! This task is particularly well suited for Humanities and Social Sciences.**

Time required: Hours. Essay writing. Can be done in parts.

Explanation:

This task concludes the use of Voyant Tools in the material package. In it, you can utilize everything you have learned so far in the previous tasks – you will not be provided hands-on, step-by-step instructions this time.

Learning objectives:

- **Learn how to do independent corpus research with Voyant Tools**
- **Figure out which options suit you the best**
- **Use Voyant Tools independently**

The task:

Decide on an issue or a concept that is central to your studies that you want to study further. Write an essay on the topic, and utilize Voyant Tools in any way you see fit to supplement your analysis (linked concepts, differences in definition, typical linguistic contexts and so on).

- Write an essay of 2-3 pages. Write in the manner that an educated person, outside your field of studies, can understand it.

Task 10: AntConc: Concordance tables and lexical bundles

- **Note! You cannot do this task on a public computer without administrator rights. You should do this on your own computer.**

Time required: 60-90 minutes. Should be finished in one session.

Explanation:

In this task, you will learn to use a different program, AntConc, which is dedicated concordance tables and the study of collocates. It is not as versatile as Voyant Tools, but is more reliable, and better for searching collocates, words that pair with another word. AntConc requires a bit more work than Voyant Tools – so a tutorial is useful.

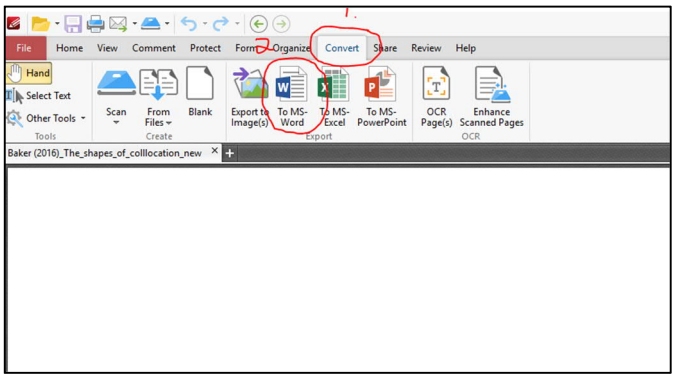
Learning objectives:

- **Become familiar with a new corpus tool: AntConc**
- **Find out about the options AntConc has: finding out the most frequent words, concordance analysis, collocations, phrase analysis**
- **Deepen your knowledge of the principles of corpus linguistics**

1. Go to http://www.laurenceanthony.net/software/antconc/ and download the version of *AntConc* supported by your operating system.
   - There are also handy Youtube tutorials available for AntConc in this page.

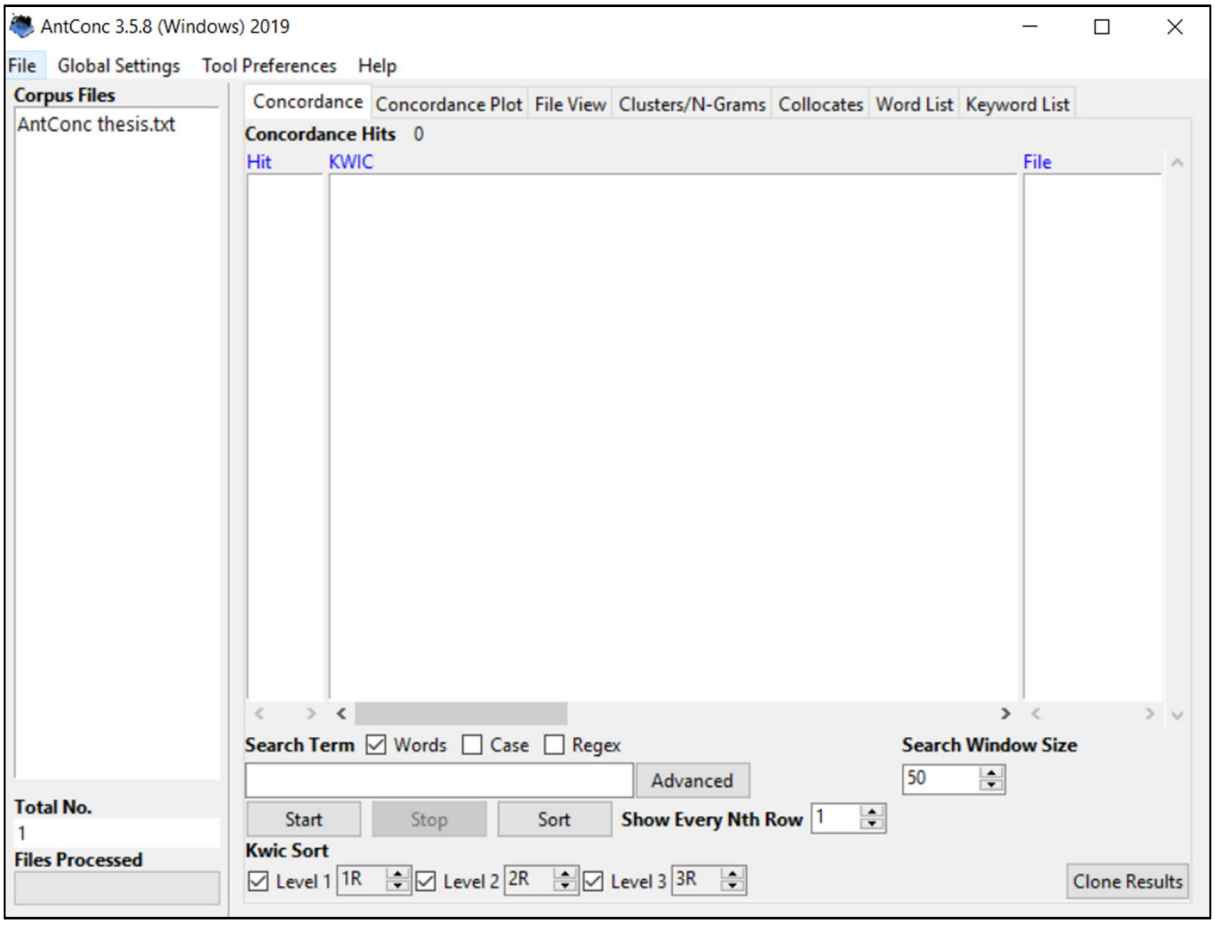
2. Install AntConc on your computer.

3. AntConc does not support pdf or word files. Therefore, you need to convert a file into .txt.
   - You can use the cleaned documents you made in Task 7 when you created a linguistic reference corpus and convert them to .txt.
   - If you haven´t done task 7 do the following:
     - First, convert a pdf file into a word file:



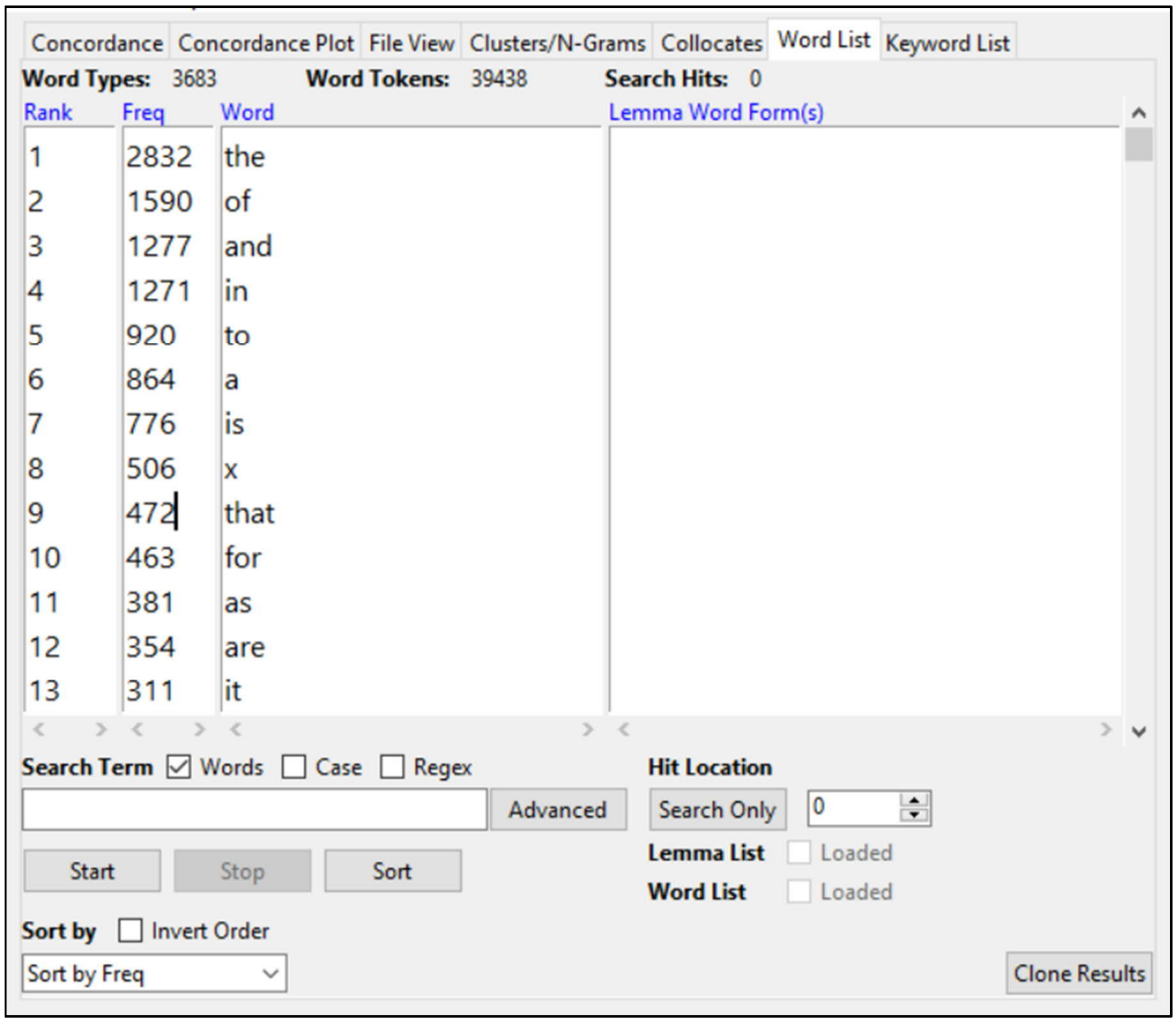
- Second, open the word file, and choose *Save as*, and select .txt from the options.

4. Now you have AntConc installed and a text file to enter to it.
   - In the example, I demonstrate how AntConc works using a .txt version of my own thesis as an example. This is the view you get:

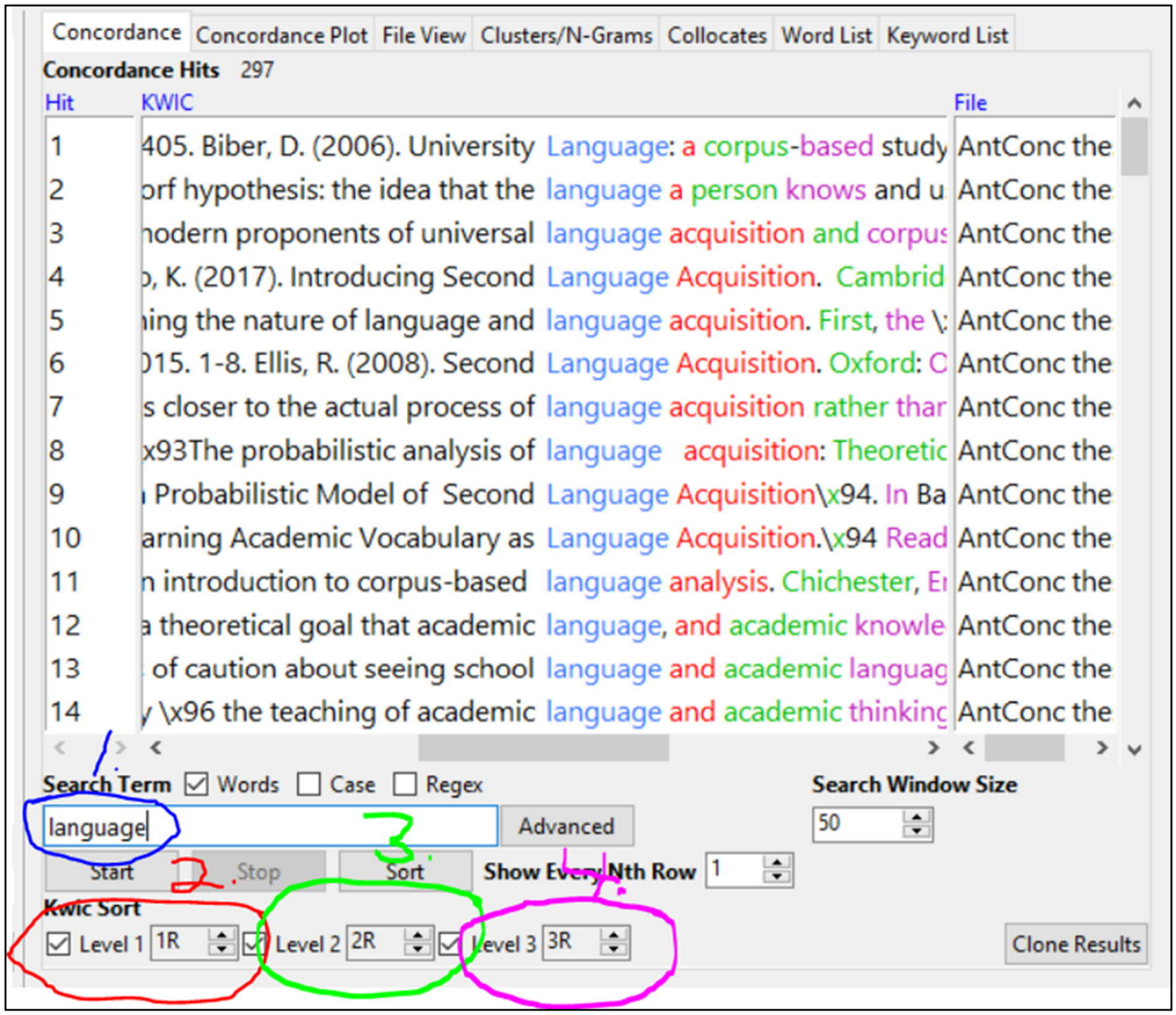


   - Open “File” in the upper left corner and choose “Open File(s)” and find the document you converted to .txt.
     - In the picture above, I have already downloaded a .txt to AntConc in the picture above.
     - You can upload multiple files to analyze a corpus, and remove any that you don´t want to analyze easily from options under “File”.

5. Choose the *Word List* option from the second highest bar. Click on it.
    - The other options that are particularly useful are *Concordances* and *Clusters/N-Grams*.



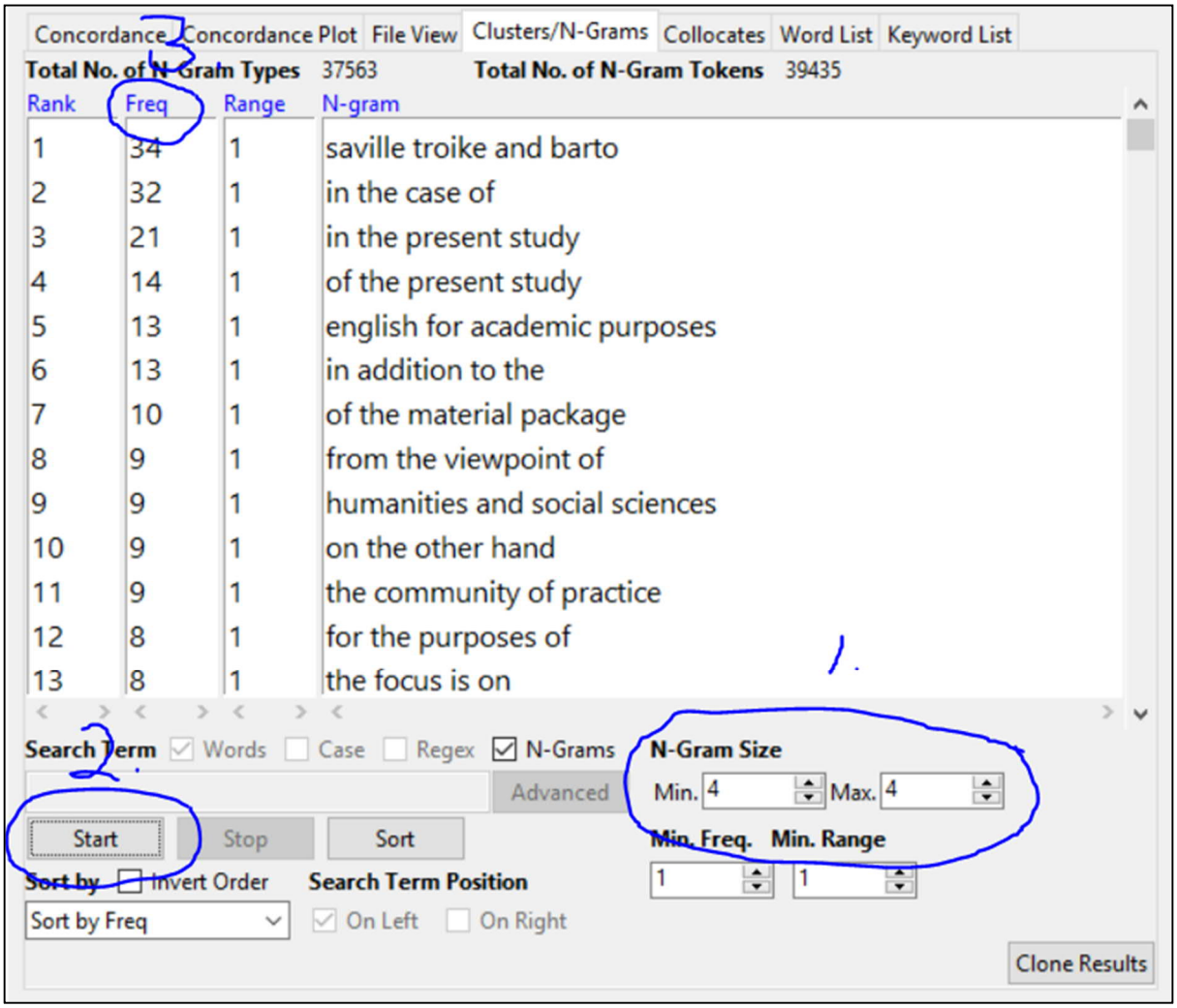
    - Note that AntConc shows *all* the words, even articles such as the, unlike Voyant Tools. It also shows non-words such as x.
    - This is how you find out which words are frequent in a document or a corpus with AntConc.
6. Next, switch to the Concordance view from the bar above.
    - Concordance tables are *the* corpus linguistic method for analyzing how words are used in context, and AntConc is better in this regard than Voyant Tools, as we shall see. In the screenshot on the next page, I have typed the word "language" as a search term – as the Key word in context, *Kwic*.



- AntConc uses color to distinguish the *Kwic* from the other text. You can search one to three collocates for the word at any direction and distance up to 20 words from the Kwic, the search term.
    - The Search Term is in Blue
    - The first collocate is in Red
    - The Second collocate is in Green
    - The third collocate is in Purple
- Note that you can specify the location for each of the collocates or turn any of them by removing the mark from the Level 1, Level 2, or Level 3 box.
- If the *Kwic* term is frequent in the corpus, the window only fits some of the *concordance lines*, the parts of sentence(s) that are around the it, and you will have to scroll every now and then to study more occurrences.'
    - Concordance analysis can be time-consuming – watch out for "concordance burnout" when you do it!

- Explore different options with at least four words and different collocation locations.
    - Going back to word list is a good option for finding words.

7. Now that you have familiarized yourself with *Concordances*, switch to *Clusters/N-Grams*.
    - N-Grams are basically combinations of words, similar to lexical bundles in the glossary.
    - This option is great for finding out lexical bundles and common expressions in your text or corpus. There are three things to take note of:



1) First, the option of specifying the range of the sizes of N-grams sought.
    - Three to five is a good option for finding lexical bundles, four is, by some definitions, the default size of a lexical bundle
    - Note that some of some of the n-grams are not lexical bundles in the sense that they serve a grammatical function in the text. For example, the most frequent one in the example is a source. So you need to do some work.

2) Second, obviously, the start button. Press this after you have specified the size.
3) Three, the frequency count of the n-grams.
    - Here, you can see how frequently a phrase occurs in the corpus.

8. Explore the different options. Two word n-grams (word pairs) and larger.

9. **You are done! You will need to use these options in the next two tasks.**

Task 11: AntConc as a linguistic reference

- **Note: You will be working on AntConc, so make sure you have it installed.**

Time required: 90-120 minutes. Can be returned to.

Explanation:
In this task, you will practice in using AntConc as a linguistic reference. Since you already know how to clean up a corpus and have cleaned up a corpus in Task 7, and practiced the basic features of AntConc in Task 10, you should not find this too difficult.

Learning objectives:

- **Learn to use AntConc as a linguistic reference**
- **Notice the types of collocations professional writers in your field use**
- **Notice the types of lexical bundles used in your field**
- **Develop your Academic English**

1. Convert the cleaned up word files into .txt files in your *corpus_building* folder.
   - Create a new folder for all the .txt files – this way you can upload them all simultaneously to AntConc.

2. Open AntConc, then open the text files in it for analysis.
   - You can use the "Open Dir" feature to open all the .txt files in the new folder.

3. You can start using your Corpus as a reference tool in writing in English. There are at least three ways you can use it for this:
   - Checking how a term is commonly used in your discipline from the concordance table. Enter the term you are wondering about as a search term, and spend some time looking at the contexts of use.
   - Checking the collocates for a word. This can be done from the concordance table.
     - For collocates that are right next to the word, the *Cluster/N-Gram* view can be used: simply enter the word as a search term and you will get the paired words when the size of the N-Gram is set as 2.

- Finding out common lexical bundles in your field from *Cluster/N-Gram* view.

4. Come up with 10 words that you were wondering about that you want to use in your writing or that are frequent in your field of studies.
   - Write them down on a word file as bullet points
   - You can use the *Word List* view if you are short on ideas.

5. Check how they are used in your corpus from the *Concordances* view.
   - Write a couple of sentences of your own where you summarize your finding
   - Write your findings under every word.

6. Check the collocates for these words – what words are they paired with?
   - Use the *Cluster/N-Gram* view and set the minimum size as 2, and the maximum size as 3 – so that clusters including an article (a, the..) are included.

7. Now, you should check the lexical bundles that are used in your field. Go to the *Cluster/N-Gram* view and set the size minimum and maximum size to four – many bundles that serve a purpose in a text are four words.
   - Do not include technical terms that are four words in length that occur in texts, if any
   - Write down the ten most common lexical bundles in your field to the word file.

8. Ok, now you have the words, and you have the lexical bundles. The final step is to practice using the lexical bundles. Come up with sentences that include one of the words from step 4, and one of the lexical bundles from step 7, and write them down.
   - Tip: some of the lexical bundles start a side sentence.
   - Tip: You can also come up with a whole text using the ingredients – but it will be more challenging!

9. **Congratulations! You have completed task eleven!**

Task 12: AntConc to improve your own texts

- **Note: You will be working on AntConc, so make sure you have it installed.**

Time required: 90-120 minutes. Can be returned to.

Explanation:

In this task, you are given ideas on how to use AntConc to check your own texts and compare them to your reference corpus, the professional writing in your own field. This task is paired with task 11, like task 2 was paired with task 1. You should have the converted, cleaned up files in your corpus by now. You also need a sample, or multiple samples of your own academic writing: course essays, thesis work, and so on.

Learning objectives:

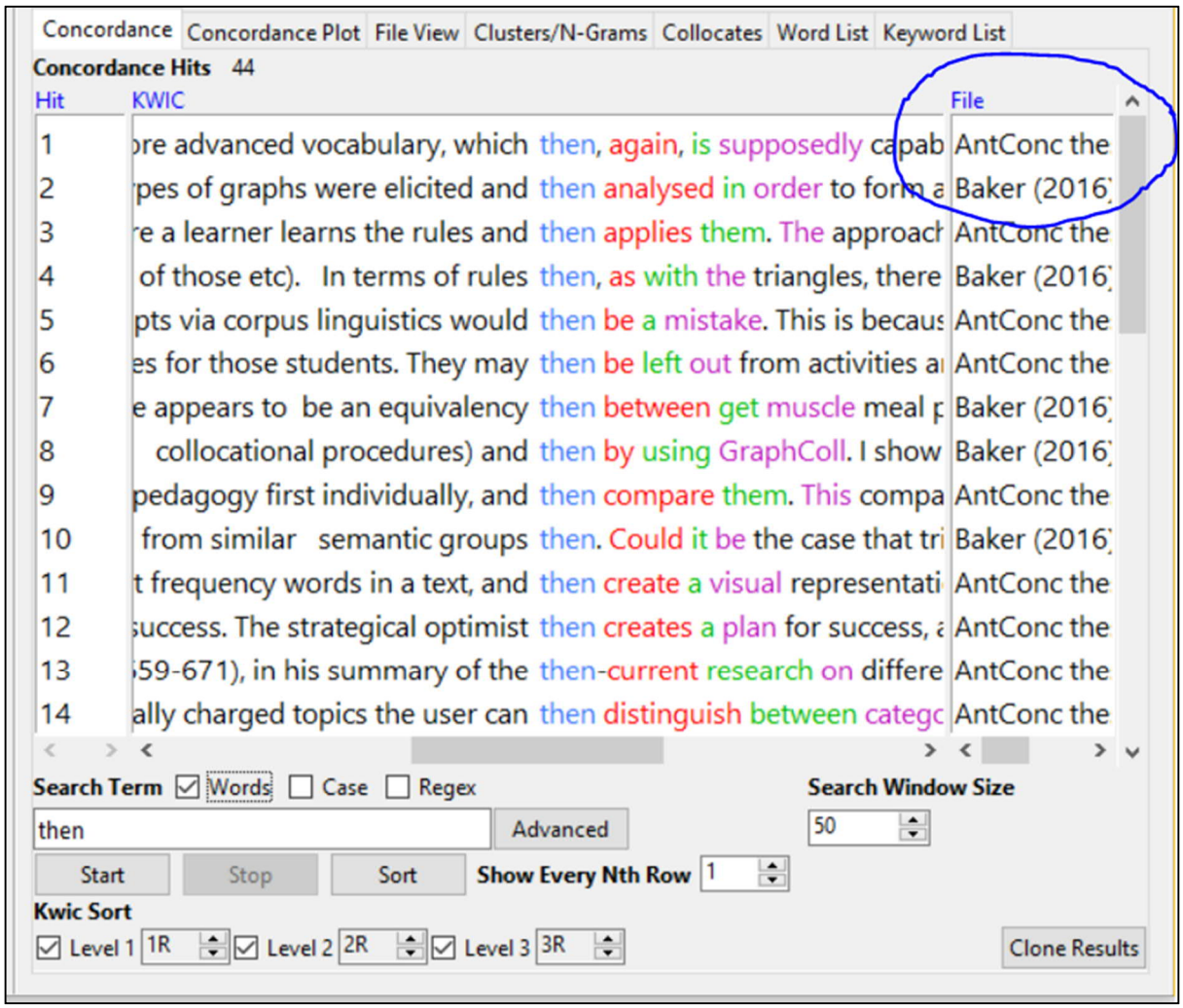
- **Grow more familiar with AntConc as a tool**
- **Learn to study your own texts with AntConc**
- **Notice areas of improvement in your writing and make it more diverse**
- **Develop your Academic English**

1. Make a corpus of your own essays for AntConc: convert them from docx file into txt.
    - When you choose the "Save as" option when converting, them, also rename them, starting with own_writing (and then the original file name)
        - This way you can tell them apart from professional writing samples more easily in the concordance table when you are simultaneously comparing the texts.

2. First load both your own writings and the corpus you have built into AntConc.

3. Check the concordance table for 10 words that you have used in your writings.
    - Remember, when you are using this on your own, you can check any word, just one, or as many as you want.
    - Now, the fact that you renamed your own writings comes handy.

- As you can see in the screenshot in the next page, the concordance table in AntConc mixes up from the key word hits between files, and shows the file in the side. They might look similar in some cases.



4. Now you can check what sort of lexical bundles are used in this ´combined corpus´ of yours. Go to the *Clusters/N-Grams* view. Select 4 as the minimum and maximum size of the N-Gram.
   - Remember that the amount and length of your own documents may be different – you are only doing this as a preparation for the next stage.

5. To get a better idea of the types of lexical bundles and phrases you use in your own writing, you need to remove some of the files in AntConc.
   - Select the files from the professional corpus in the corpus file bar left to the concordance table. Then choose “Close selected File(s)” under *File*.

6. Now, go the *Clusters/N-Grams* view. Choose minimum of 4 and maximum of 4 for the size of N-Gram.
   - Mark down the lexical bundles you used on a word file.
     - Mark at least 20 lexical bundles.

7. Now, reopen your own professional corpus on AntConc, and remove your own essays.
   - This should be easy because of step 5.

8. Stay with Clusters/N-Grams view. Choose minimum of 4 and maximum of 4 for the size of N-Gram.
   - Mark down 20 lexical bundles to the same word file you marked your own lexical bundles to, but make sure they are separate.

9. Compare the lexical bundles you have been using with the ones in professional corpus,
   - Notice similarities and differences.

10. When you have time, read some of the texts and pay extra attention to how the professional authors use these bundles.
    - What sort of function they serve in the text?

11. **You are done! You have learned how to use AntConc to compare your own writing with professional writing.**

You can be creative with both Voyant Tools and AntConc now – explore the options and come up with creative ways of using them!

Adaptation Ideas for Classroom

Even though this is a self-study package, teachers may want to utilize the tasks presented in it for classroom work.

Some of the tasks work well as pair/group work tasks:

- Tasks 1-3 may include discussion on the vocabulary and concepts
- Tasks 4-6 can include comparison of writing norms displayed in each students corpus
    - This is particularly fruitful in an EAP course with students from different disciplines attending
- Task 7 may include collaborative work in cleaning and building a corpus with students from the same discipline: they can share the documents they have cleaned up with others to reduce the workload and to build a bigger corpus-
- Task 9 may include peer feedback or review of the essays, and possibly a presentation of the findings
- Tasks 10-12 can be done in a classroom setting to provide support for the less technically savvy students – provided the classroom has computers where AntConc is available.

In addition, some of the tasks can be used for students in secondary education or even primary.

- Task 1 works even in primary education with different instructions
- Tasks 2-6 might work in high school
- Tasks 7-12 require more support from the teacher if they are used in secondary education, or the requirements (essay length etc.) need to be reduced.

The Corpus linguistic tools for other text types

The tools and tasks presented in the material package can be adapted to study other texts, not just academic articles.

- Any of the tasks work well with newspaper articles
    - You can find how ideologically charged words (terrorism, nationalism, etc.) are used in a newspaper articles by building a corpus of them
- Fiction writers' most typically used words in an interesting field of study
    - This is the default option in Voyant Tools: Shakespeare and Austen
- Teachers can make a combined corpus of learner texts such as essays
    - They can use corpus tools to identify common errors.
- … And there are many more uses.