Second language lexis and the idiom principle

Svetlana Vetchinnikova

Academic dissertation to be publicly discussed, by due permission of the Faculty of Arts at the University of Helsinki, in auditorium XIII, University main building, on the 29th of August 2014, at 12 o'clock.

Department of Modern Languages
University of Helsinki

Cover illustration: Klaudia Rastorgueva

ISBN 978-951-51-0063-4 (paperback)
ISBN 978-951-51-0064-1 (PDF)
http://ethesis.helsinki.fi

Unigrafia
Helsinki 2014

Abstract

This work sets out to examine how second language (L2) users of English acquire, use and process lexical items. For this purpose three types of data were collected from five non-native students of the University of Helsinki. First, each student's drafts of Master's thesis chapters written over a period of time were compiled into a language usage corpus. Second, academic publications a student referred to in her thesis were compiled into a corpus representing her language exposure. Third, several hundreds of words a student used in her thesis were presented to her as stimuli in word association tasks to obtain psycholinguistic data on the representation of the patterns in the mind. Lexical usage patterns, conceived of in accordance with John Sinclair's conceptualisation of lexis and meaning, were then compared to (1) language exposure and (2) word association responses.

The results of this triangulation show that, contrary to mainstream thinking in SLA, language production on the idiom principle, i.e. by retrieving holistic patterns glued by syntagmatic association rather than constructing them word by word, is available to L2 users to a much larger degree than is often claimed. More than half of significant multi-word units used by the students also occur in the language they were exposed to. The 'idiosyncratic' multi-word units are often a result of approximation or fixing. Approximation is a process through which a more or less fixed pattern loosens and becomes variable on the semantic or grammatical axis due to frequency effects and the properties of human memory. Fixing, on the other hand, is a reverse process making the wording of the pattern become 'overly' fixed through repeated usage. Neither of the processes damage the meaning communicated in any way. Word association responses also support the main conclusion of the availability of the idiom principle showing that multi-word units used are also represented holistically in the mind and so confirming the continuity between exposure, usage and psycholinguistic representation. Furthermore, they suggest that the model of a unit of meaning developed by Sinclair has psycholinguistic reality as representations of lexical items in the mind seem to mirror the components of a unit of meaning: collocation, colligation and semantic preference.

This work offers an in-depth discussion of Sinclair's conceptualisation of meaning and a novel methodology for studying units of meaning in L2 use both quantitatively and qualitatively by triangulating usage, exposure and word association data. It is hoped that the dissertation will be of interest to scholars specialising in second language acquisition and use, English as a lingua franca, phraseological view of language and corpus linguistic methodology.

Contents

Acknowledgements

As the phrase goes, this dissertation would not have been possible without the help and support of many people. It strikes one as being so very true. And small wonder, after all our phraseology is polished by generations of language use and no academic work is a sole achievement of the author. I wish to express my warmest thanks to my teachers, colleagues, friends, family and, of course, the research community I consider myself privileged to be part of: everybody who has been so kind, encouraging and generous in giving their advice and expert judgment when sought.

I would like to thank my pre-examiners, Prof. Susan Hunston and Prof. Nick Ellis, who have kindly shared their thoughts with me and whose opinion is so important for me. Their excellent comments have stirred new ideas and enthusiasm for further research.

I am deeply indebted and eternally grateful to my dear supervisors, Prof. Anna Mauranen and Dr. Tuula Lehtonen. Anna, a brilliant scholar and intellectual, is also an excellent supervisor: with her extensive knowledge in a wide variety of research fields she is able to follow her student along any path one is given all the freedom to choose. And yet, being a vigilant academic who demands quality, she will not miss a single slip in argumentation. Tuula, a caring and understanding mentor, has always managed to keep a watchful eye on the practical side of things urging me to think how all my theorising applies to actual language learners and language teaching. Thank you so much for having faith in me and being there for me throughout these years.

I have had the pleasure to be part of Anna Mauranen's English as a lingua franca research group at the University of Helsinki, whose members deserve my sincere gratitude. I would like to thank Anna Solin, Jaana Suviniitty, Diane Pilkinton-Pihko, Henrik Hakala, Jani Ahtiainen, Netta Hirvensalo, Kaisa Pietikäinen, Elina Ranta, Ruut Kosonen as well as many other visiting scholars and students for years of fruitful and stimulating seminars. Special thanks go to Niina Hynninen who was my roommate for the first years after our move to Finland and who virtually led me into the life at University by the hand. Her friendship made those first steps as a 'newborn academic' so much more enjoyable. Ray Carey – colleague and friend – has always been able to listen to my animated talking on scientific as well as other topics. Our lively discussions have greatly helped me to develop my thinking. Thank you, Ray.

I am immensely grateful to the friendly and welcoming community of the Language Centre teachers for the opportunity they have given me to present my work and to put my

teaching skills to test as well as for many useful comments and observations on my work. In particular, I would like to thank Paul Graves, Mícheál Briody, Robert Moncrief and Roy Siddall for their invaluable help in the data collection. Heartfelt thanks are due to the students who agreed to participate in this research – it would really not have been possible without you.

Along the way, many people have given me first-class advice at the time when it was most needed. Laura Winther Balling was kind enough to listen to my ideas at the very early stages of research and discuss the psycholinguistic underpinnings of it with me. Jukka Tyrkkö, my teacher of Corpus Linguistics, Antti Arppe and Peter Uhrig helped to solve the statistical puzzles. Also, it was in fact Jukka who first told me about ConcGram which later became indispensable for the study. Kari Pitkänen's constructive feedback on the presentation I gave at the Language Centre was so effective that it made me think of starting to collect another type of data, which later became my C2 corpora.

I have greatly benefited from being a member of the GlobE and ChangE consortiums, Langnet doctoral programme in language studies, NordLing network of graduate schools, Lingua Francas and Plurilingualism research group. I also wish to thank the Department of Modern Languages and all its inhabitants for being a great place to be in.

I owe special thanks for the financial support to the Finnish Cultural Foundation, University of Helsinki Funds, and the GlobE consortium. I was also happy to receive travel grants from the English unit of the Department of Modern languages, the Chancellor of the University of Helsinki, Langnet doctoral programme and NordLing network. These grants enabled exciting and highly beneficial trips to conferences, symposiums and doctoral workshops, which were always a great stimulus to move the work on this dissertation forward.

On a more personal note, I wish to thank my family who have always followed me wherever I go. My husband is the epitome of patience. And nobody has ever taken care of me better than he does. I have everything I might need even before I start to think about it. I am sure the last months of my work on the dissertation have put a serious strain on him. However, I noticed that my absence from the kitchen had a remarkable effect on his cooking. Maybe I should start writing another book… My precious daughter, Mayya, was the main inspiration for me to make the decisive moves and start working for a doctorate. She is the perfect child who somehow manages to stay calm and level-headed despite her mother's restlessness.

What a pity I cannot mention everybody here who has been so important for me during these years, but I am sure they know I feel deepest gratitude for their mere presence in my life.

Helsinki, July 2014

1. Introduction

What does it mean to know a word? We rarely ask this question when acquiring our first language, but it takes on a new significance once we step into the second language (L2) territory. In Applied Linguistics, vocabulary researchers distinguish between different aspects of productive and receptive word knowledge. A comprehensive vocabulary knowledge framework (Nation 2001) includes knowledge of a word's form, meaning and use, each further subdivided into more specific kinds of knowledge. To know a word's form is to know its spelling, phonology and morphology. To know its meaning is to know its form-meaning mapping, its concepts and referents as well as its paradigmatic associations. To know its use is to know its grammatical functions, collocational associations and constrains on use, or where, when and how often the word is used (Nation 2001: 27). In addition to different aspects of knowing a word, there are also degrees of knowing it: for example, from vague to precise (Paribakht and Wesche 1993, Vocabulary Knowledge Scale). Schmitt (1998a) reported that it took him two hours to interview four students on four aspects of eleven words. In half an hour a non-native English lecturer produces 3600 word tokens (ELFA corpus). It is unlikely that we are employing all the aspects of our declarative word knowledge in language use.

Then, what does it mean to be able to use a word? In addition to the different aspects of knowledge a word requires, it often has more than one meaning. Sinclair took a simple sentence "The cat sat on the mat" and counted all the possible combinations of meanings it must generate based on the number of meanings each word in the sentence has. *Cat* has 24 meanings, *mat* – 17, *on* – 25, *sit* - 18, *the* – 15: as a result one must be working through 41,310,000 possible meaning combinations to arrive at the only correct one (Sinclair 2004 [1998]: 137-138). This casts doubt on the plausibility of independent lexical choice and suggests that words are nor produced or perceived one at a time but in association with the surrounding text. The properties of this association may shed light on the ability to use lexis.

What does it mean to learn a word? *Oxford English Dictionary* includes full entries for 171,476 words in current use.[1] An educated native speaker is estimated to have a vocabulary size in the range between 16,000 and 20,000 word families (Schmitt 2010). A language learner needs to know at least 98% of running words in order to understand a text (Hu and Nation 2000), which means 9,000 word families if it is a novel (Nation 2006). These

[1] http://www.oxforddictionaries.com

are not the numbers of words one can feasibly acquire through explicit instruction and learning. Most of the learning must happen implicitly through exposure.

Facts like these encourage a shift of focus to (1) implicit rather than explicit lexical knowledge, (2) lexical patterns or multi-word units (MWUs) rather than single words, (3) usage-based acquisition rather than explicit instruction. When applied to L2 learning, these three foci converge on L2 implicit acquisition of MWUs through exposure. This topic has generated much interest in the recent years. However, it remains unclear what MWUs are and to what extent L2 learners can acquire them from exposure and use in their own production.

There are numerous descriptions of MWUs in linguistic theory. Granger and Paquot (2008) perceptively distinguish two major approaches to phraseological patterning. The first, termed "phraseological" (after Nesselhauf 2004), traces its roots to the East European tradition and is characterised by top-down identification and classification of phraseological units on the basis of their linguistic features such as fixedness and semantic non-compositionality. It is typical in this approach to place phraseological units on a continuum from free combinations to figurative idioms (Cowie 1981). The second approach, which Granger and Paquot call "distributional" (Evert 2004) or "frequency-based" (Nesselhauf 2004), but which is also sometimes termed "corpus-driven" (Tognini-Bonelli 2001), stems from John Sinclair's corpus linguistic work in lexicography and builds on automatic extraction of co-occurring and recurring items from text. The more recent approach has uncovered the pervasive nature of regularities in text and "pushed the boundary that roughly demarcates the 'phraseological' more and more into the zone previously thought of as free" (Cowie 1998). Indeed, by declaring "[t]he phrase, the whole phrase and nothing but the phrase" (Sinclair 2008: 407), Sinclair puts phraseological patterning forward as a characteristic property of language as a whole.[2] This view is distinct from seeing collocational associations as an aspect of word knowledge or making allowance for the existence of a stock of phrases in addition to a stock of words, from which items can be drawn.

Meanwhile, in SLA the attention of the scholars is captured by the problems L2 learners and users[3] seem to have in acquisition and use of MWUs. We hear that L2 learners suffer from "collocational dysfunction" (Howarth 1998: 180), that their "phraseological skills are severely limited" (Granger 1998: 158) and that "the non-native speaker, however accurate

[2] Cf. Ellis (2012b): "language learning is, in essence, the learning of formulaic sequences and their interpretations" (17).

[3] The distinction will be explained Chapter 3. See Mauranen 2011 for an extensive treatment of the question.

in grammar and knowledgeable at the level of words, would always be a potential victim of that lesser store of formulaic sequences" (Wray 2002: 210). It is considered that phraseological competence is hinged on the ability to acquire, store and retrieve MWUs holistically from memory, which appears to be compromised in the case of L2 learning. In Sinclair's terms, while native speakers (NSs) predominantly operate on the idiom principle, non-native speakers (NNSs) are apparently forced to rely on the open-choice principle (Granger 1998; Seidlhofer 2009; Wray 2002). These observations form the major impetus for the present study.

1.1. Research data and questions

Five non-native English students from the University of Helsinki participated in this study. To examine phraseological competence of these L2 users, three types of data were collected from each of them: a corpus of Master's thesis drafts they were writing in English, a corpus of academic publications cited in the thesis and a database of word associations elicited in response to stimulus words from the thesis. These kinds of data were taken to represent each student's language usage, priming language and psycholinguistic associations. The research questions are twofold: on the one hand, they probe the availability of the idiom principle to L2 users and, on the other, the psycholinguistic reality of Sinclair's model of a multi-word unit. These two main issues subdivide into more specific questions:

1. To what extent is the idiom principle available to L2 users?
    (1) Do L2 users acquire units of meaning implicitly through exposure?
    (2) Do L2 users operate with units of meaning in language production?
    (3) Is there evidence of psycholinguistic representation of the units of meaning attested in L2 production?
2. Is the model of a unit of meaning psycholinguistically real?
    (1) Are the components of a unit of meaning psycholinguistically associated?
    (2) What further properties does syntagmatic association exhibit?

These research questions are addressed by comparing language usage data (1) to the priming language and (2) to the word association data.

1.2. Structure of the thesis

This thesis is organised in seven chapters. Chapter 2 lays down the theoretical framework of the study. It introduces a corpus linguistic approach to language patterning and concentrates on Sinclair's conceptualisation of lexis and meaning zooming in on the concepts of collocation, unit of meaning, semantic prosody, the idiom principle, delexicalisation and meaning-shift. The conceptual system arising from this theoretical analysis not only guides the empirical research but also informs the interpretation of L2 phraseological competence presented in Chapter 3. Chapter 3 reviews and revisits mainstream research into L2 acquisition and use of MWUs, it analyses the possible reasons behind the common conclusions about the deficiency of L2 phraseological competence, offers an alternative explanation and suggests a cognitive underpinning for it. Chapter 4 presents the methodological decisions of this study which are essentially non-orthodox. It explains the three types of data collected, dwells in particular on the word association method, discussing the history of its application to research with an attempt to develop a better understanding of what it actually taps and what lessons can be drawn with regard to its design and administration. Chapter 4 also looks at the structure of the study and the basic principles of analysis.

The empirical work in this study is divided into two parts. First, the usage patterns are compared to the priming language. This part of the work is presented in Chapter 5. Then, the usage patterns are compared to word association responses. Chapter 6 takes care of this second part of the work. Both chapters contain qualitative as well as quantitative analysis of the patterns arising from the comparisons. Chapter 7 summarises the findings first spelled out in the respective chapters and integrates them into the models of a unit of meaning and the process of meaning-shift discussed in Chapter 2. The proposals put forward in Chapter 3 are also taken into account in this modelling. Each chapter is supplied with an introduction giving more specific guidance on the contents and the line of argument pursued.

2. Unit of meaning and the idiom principle

This chapter presents the theoretical framework on which the study is built. It is based on an interpretation of Sinclair's conceptualisation of lexis and meaning. The argument for this interpretation is made in detail, and, Sinclair's conceptualisation is discussed step by step with special attention given to debatable concepts such as collocation, which enjoys a whole number of different definitions, and semantic prosody, whose controversial nature provokes book-length treatments.

A number of linguistic theories and approaches to lexical patterning come close to the framework advocated in this study. Not all of them are discussed for reasons of space, but only those deemed to be most relevant. Lexical priming theory, Louw's semantic prosody, as well as Wray's formulaic language are given separate sections at the end of the chapter, many other approaches are discussed in conjunction with specific aspects of Sinclair's conceptualisation. So in this chapter I will first look at different perspectives on the phraseological phenomenon, then move on to discuss Sinclair's proposals and in the end compare Sinclair's views with other approaches.

2.1. Phraseology: an anomaly or a characteristic property of language?

The phraseological phenomenon is described with an impressive variety of terms (see e.g. Wray 2002: 9): chunks, clichés, routines, fixed expressions, multiword units, fossilised forms, unanalysed chunks, lexical phrases, irregular phrases, formulaic sequences, collocations, to name but a few. It is indeed disputable whether these terms focus on one and the same phenomenon, but what seems to be common for all of them is that they emphasize the special status of some linguistic items/units. As Wray points out in her oft-cited work on formulaic language, "if there is a standard view of what formulaic language is [...] at its heart will be something about word strings which 'break the rules'" (Wray 2002: 261). In this view phraseology is an anomaly in an otherwise rational language. However, the picture looks very different once we start to realise that the patterns we have been able to identify so far are only the peak of an iceberg. The more fixed a multi-word item is, the easier it is to detect it as 'anomalous': we notice that the whole item consisting of more than one word or some part of it recurs verbatim and are thus able to pinpoint its boundaries or we calculate that the item as a whole means something different form merely the sum of the words it consists of. The matter becomes much more complicated when there is no verbatim repetition or drastic change in meaning. With corpus linguistic methodology, it has become clear that phraseological patterning is much more pervasive than we were able to imagine and apparently reveals a general property of a language rather than an anomaly, the tendency for “syntagmatic organization in language in use” (Stubbs 2009:115).

Corpus Linguistics has made possible to observe language in a way that makes visible the patterns which are otherwise not discernible for human analytic abilities. Michael Stubbs (2011) in his plenary lecture at ICAME 32 drew an illuminating parallel between the kind of observation Corpus Linguistics enables and the kind of observation that led Darwin to his theory of species. Apparently, a drawing of finches from the Galápagos Islands where they are presented in a convenient tabular way – one under another and facing the same direction, just like ordered concordance lines – helped Darwin to see that in spite of certain undeniable differences, the birds represented one and the same bird family and the differences are the consequences of natural selection and evolutionary change. In the same way, the concordance view Corpus Linguistics offers is able to highlight not only the differences but also the similarities in patterning, leading us to a conclusion that a whole number of word sequences are in fact instances of one pattern.

2.2. Unit of meaning: the model

One thing that corpus linguistic observations of language patterning suggest quite clearly is that an orthographic word should not be considered a unit of meaning by default, in other words meaning does not necessarily or even normally reside in a single word. Therefore, a lexical model based on orthographic words is extremely unhelpful: it “claims more meaning in an expression than is actually usable” (Sinclair 2004 [1998]: 140) due to syntagmatic constraints.

Finding a reliable form-meaning pairing is a challenging task. When analysing a stretch of text, a researcher is aware of the meanings expressed there, but the forms with which these meanings are expressed remain to be individual instances on the basis of which it is not possible to draw conclusions about the common forms these meanings can take. Sinclair calls these forms ‘canonical’ and postulates that for each lexical item it should be possible to find one canonical form with all the rest of its instantiations regarded as its variants (Sinclair et al. 2004: xxiv, the OSTI report originally published in 1970). In contrast, when observing concordance lines, the forms become clear, but then a researcher loses sight of the meanings expressed. For this purpose, it would be necessary to go back to the context of each identified form, but how large should the context be? Very often the context of a concordance line is not enough to draw conclusions about the meaning. Even if we take the whole text into account, this would still leave out a lot of aspects such as the context of the text and intertextuality, yet making analysis of concordances an impossible task.

Sinclair compares the problem of relating syntagmatic and paradigmatic axes of meaning with Heisenberg's uncertainty principle in atomic physics: just like an atom whose position and momentum are not simultaneously observable, a word's meaning can be described either from the point of view of syntagmatic axis or solely paradigmatic axis, but it is hard to take into account both at the same time (see Sinclair 2004 [1998]: 141). That is, a word which is assumed to be the main bearer of meaning is usually studied either in the context of one text, which leads to its paradigmatic discussion, i.e. what a particular word means in this particular stretch of text and how it can be substituted, or across texts, which reveals its co-occurrences. Syntagmatic or horizontal observation of a word, i.e. in context, allows drawing conclusions as regards its paradigmatic capacity. Paradigmatic or vertical observation of a word, i.e. across texts, allows observing its syntagmatic behaviour.

Sinclair's model of a unit of meaning is a solution to the problem of incorporating the information from both axes in a form-meaning pairing. In his model Sinclair breaks away from the idea of an orthographic word as a major building block, and instead talks about fixed obligatory components, the core and semantic prosody, and optional variable components, collocation, colligation and semantic preference. The core of a unit does not have to be represented by a word or a certain number of words. Instead, it is defined as the most invariable form which can be identified for the unit. Likewise, semantic prosody is the most uniform meaning of a unit as a whole, i.e. the meaning which is always realized no matter which other components are participating. This means that the core and the semantic prosody form the nucleus of a form-meaning pairing in a unit. The optional components which allow for internal variability are both the result of normal linguistic variation and the mechanism enabling the unit to adapt to specific contexts. The model incorporates a possibility for a specific co-occurrence relationship, i.e. a verbatim association (collocation) and more abstract associations: with a grammatical feature (colligation) and a semantic feature (semantic preference). In other words, colligation allows for variability within a grammatical class, while semantic preference tolerates variation within a semantic set. The fact that these components are optional means that they may or may not be realiszed, most importantly they allow for paradigmatic subsidiary choices within a syntagmatic model.

To illustrate how the model works, I will now use one of Sinclair's examples: a unit of meaning *naked eye* (Sinclair 1996a). Table 2.1 presents the full extended unit divided into components according to the model. The phrase *naked eye* itself is a collocation because it is a verbatim co-occurrence of two words. In Table 2.1 it is split into an *origin* and a *co-occurring word*: the terms suggested in Cheng et al. 2009 in conjunction with analysing the structure of 'concgrams', or co-occurring words regardless of positional or constituency variation (Cheng et al. 2006; see also Sections 2.7 and 4.2.2). 'Origin' is a corpus query search word; 'co-occurring word' is the one which the corpus query shows to be co-occurring with the origin. But in our case *naked* could just as well be the origin, and *eye* – the co-occurring word.[4]

Table 2.1 Unit of meaning *naked eye*[5]

	semantic preference	**colligation**	**collocation**			
			co-occurring word	co-occurring word	origin	
It is not/barely *It cannot be*	*visible* *obvious* *discernible* *spotted* *seen*	*by* *with* *via*	*the*	*naked*	*eye*	etc.

The co-occurrence with the definite article *the* is also important. It would be reasonable to suggest that this is a colligation because *the* is a grammatical word and has little lexical meaning. Yet, I would argue that since *the* cannot be replaced by an indefinite article in this context and is therefore invariable, it is a verbatim association and therefore a collocation. Sinclair himself includes *the* into the core together with *naked eye* since *the naked eye* forms an almost invariable sequence. My argument would be that even if variability is not permissible, there is still a chance that it will be introduced. In that case, the concept of ~~a~~ collocation defined as a verbatim association between words leaves a possibility for this departure from the established phrase if the association is loosened. For example, it turns out that *unaided eye* occurs in the BNC 7 times with exactly the same co-text as n*aked eye,* including the co-occurrence with the definite article. So though *the unaided eye* is not

[4] At the same time it must be mentioned that collocational associations are often asymmetric, that is, the probability of word A co-occurring with word B is different from the probability of word B co-occurring with word A. This seems to be true both for corpus based probabilities and for human cognitive associations (see Michelbacher et al. 2011 and the discussion in Section 6.3.5).

[5] Table 2.1 does not give an exhaustive account of all the specific co-occurrences of the *naked eye* that can be found in corpora or were described by Sinclair: it is only intended as a summary of the main argument.

mentioned in dictionaries (at least not in the Oxford dictionaries) on a par with *the naked eye*, it exists. But even this is not the point. The point is that even if either *naked* or *eye* was replaced with a close synonym by mistake, the unit of meaning would still be recognisable.[6] Therefore, *the naked eye* cannot be the core if it has to represent an invariable formal component of the unit.

Going back to the table and the components of the unit, we note that *naked eye* colligates with the class of prepositions, that is, it co-occurs not only with one specific preposition e.g. *by*, but with several, all of which belong to a grammatical class of prepositions.[7] And finally it has a semantic preference for a semantic set of 'visibility', i.e. it does not co-occur with a specific word e.g. *visible* but with a set of words which can be grouped on the basis of their semantic properties: all the words listed in the column "semantic preference", both adjectives and verbs, have something to do with the ability to see.

Still, the observational problem raised above remains: it is possible that the form chosen as the origin for query generates more than one meaning in practice, and therefore, the conclusion should be that it participates in more than one unit of meaning. It is also possible that this form is just a part of a longer unit of meaning, if the analysis shows that a longer stretch of text correlates with a constant meaning. This means that in order to arrive at the canonical form of this unit of meaning whose actual instances of occurrence vary slightly from this form but inside the postulated boundaries, it is necessary to go through each occurrence and analyse the meaning expressed in each case. The stability of the unit is ensured by the fact that in roughly all of the instances the unit was used to express the meaning that 'something was difficult to see'. This invariable meaning which is always realised whenever the unit is employed is the semantic prosody of the unit. Semantic prosody

[6] Later it will be proposed that such a replacement constitutes a mechanism behind approximation typical of second language users (see Sections 3.5 and 3.6).

[7] In this particular example, the prepositions participating in the unit of meaning are determined by different factors. On the one hand, *the naked eye* as an "instrument used to perform an action" (*Oxford Dictionary of English* 2010, '*with'*) takes the preposition *with*. Or, it also combines well with the preposition *via* implying "by means of" (*Oxford Dictionary of English* 2010, *'via'*). On the other hand, its co-occurring words from the category of semantic preference often govern the prepositions which follow. For example, *seen* is often used with *by (seen by the naked eye)* in addition to *with*, as if the naked eye was an "agent performing an action" (*Oxford Dictionary of English* 2010). Adjectives such as *visible, obvious, evident, discernible* require the preposition *to* instead: in fact according to *Collins COBUILD Grammar Patterns* (Francis et al. 1998), they share the 'Recognizable' and Obvious' groups of the pattern "ADJ *to* n" (470). Yet, the fact that the prepositions used with *the naked eye* may be governed by verbs and adjectives preceding it is not in conflict with modelling the unit as a unit of meaning and subsuming the prepositions used under the category of colligation. Through proximity the prepositions which are determined by the co-occurring words may come to be associated with *the naked eye* itself. While I would argue that a unit of meaning is the smallest independent lexical item which has relatively complete meaning of its own, it is not the smallest unit or the only linguistic unit: other units might be embedded, overlapping or bordering with units of meaning, like the Grammar Patterns, which gives rise to a complex interaction between them.

is a functional meaning of the unit as a whole: we select this unit of meaning co-selecting all its components through syntagmatic association first and foremost because we want to say that something is difficult to see. The concept of semantic prosody is not free from controversy and will be discussed in detail in Section 2.6.

This is how Sinclair's model unites both the syntagmatic and paradigmatic axes. The model is in itself syntagmatic: those items which are co-selected are allowed inside. In other words, its components are glued together by syntagmatic association. However, colligation and semantic preference, these approximated associations, allow for paradigmatic variation inside this syntagmatic model. In such a way both axes of meaning are combined in one model. More importantly it is a combination of both paradigmatic and syntagmatic axes within one unit of meaning which stretches our understanding of meaning. In his 2001 book, Michael Stubbs points to a very interesting aspect of Sinclair's model he develops further - lexical relations which are included in the model "correspond to the classic distinctions between syntactics, semantics and pragmatics, which were drawn by Morris in the 1930s (Morris 1938)" and, therefore, the model "brings lexis fully within the traditional concerns of linguistic theory" (Stubbs 2001: 88-89). This is exactly what the model does; however, in Sinclair's model the relations exist *within one meaning* and not between meanings.

2.3. Single-word units

A model of a unit of meaning does not exclude the possibility for a single word to be a unit of meaning. The optionality of collocation, colligation and semantic preference implies that a unit of meaning can consist of the core and the semantic prosody only. Therefore, when a single word, the core, has an independent meaning of its own, it can function as a unit of meaning. Its independence would mean that it can be used alone, i.e. without requiring the presence of other words, to express a particular communicative purpose – the semantic prosody. Examples are e.g. modal adjuncts (*presumably, obviously*), connective adjuncts (*however, therefore, hence, moreover*), evaluative adjuncts (*fortunately, ironically*,) conjunctions (*but, and*).

A look at the concordance of *presumably* shows that this item forms no obvious patterns of use. Seemingly, it does not collocate with any other items, does not enter into colligations and does not have semantic preferences. The only decipherable tendency for it is to appear at the beginning of a sentence and to be separated with commas, parentheses or a dash (see Table 2.2). Therefore, arguably the item is able to make meaning on its own and constitutes a separate lexical choice of a speaker. *Presumably* is often categorized as a hedge (e.g. Hyland 2005; Carter and McCarthy 2006) and this would be its functional meaning: the item appears in the discourse when the speaker chooses to be tentative.

Table 2.2 Co-occurrence patterns of *presumably*

	Origin, the core	
,	*presumably*	,
(		
-		

This is of course not a new way of looking at the word. *Presumably* is usually treated as a sentence adjunct. In Huddleston and Pullum's Grammar (2002), *presumably* is categorized as a modal adjunct, a "quasi-strong" modal adverb in the four level system, "between the strong of necessary and the medium of probably". It is also grouped with *apparently* and *seemingly* on the basis of the meaning they convey: all of them "suggest a qualified acceptance of the proposition" (Mittwoch, Huddleston and Collins 2002: 769). In Pattern Grammar, we read: "*presumably* is often found at the beginning of a sentence or clause, where it serves to comment on the whole clause" (Hunston and Francis 2000: 43). So the current analysis does not question the traditional understanding of the function of the word *presumably*, but draws attention to the fact that it has a functional meaning on its own, without the contribution of other words, unlike e.g. *naked eye* which requires all the other components in order to convey the functional meaning 'it is difficult to see'.

A single-word unit is thus a structurally possible representation of the model: it is a type of units consisting only of an invariable part. A chemical element comes to mind as a comparison: if a unit of meaning is a chemical element where its fixed semantic prosody is responsible for its stable 'chemical properties', then different realisations of the form (since each of the optional components may be realized or not) are isotopes of a unit.

A single-word unit of meaning is a limiting case of the model, but it is extremely important for the conception of lexis. It means that if we take meaning as a starting point for our approach to lexis, there is no dividing line between single words and multi-word units. Instead, the dividing line goes between lexical items with incomplete meaning which is dependent on the surrounding co-text and 'independent' lexical items which can function on their own.[8] In other words, a unit of meaning is an independent lexical item. Natural language is comprised of independent lexical items. We are faced with an incomplete lexical item when it is taken out of context with insufficient co-text. Many high-frequent lexical words of English, especially verbs like *take* or *make*, are often needlessly isolated from their habitual co-text and analysed as independent lexical items which they rarely are. This line of reasoning leads us to the concepts of core meaning, delexicalisation and meaning-shift which will be examined in the following section in the light of comparing Sinclair's and Firth's approaches to lexis and meaning. But before that, several terms, namely a word, a lexical item and a unit of meaning, are in need of a little clarification.

A word is used in this study in its purely orthographical sense as "a string of characters lying between spaces" (Sinclair 2004 [1998]: 131). A lexical item is used as a generic term and can be applied to any item which has lexical meaning. In this sense a word is a lexical item, but a unit of meaning is a lexical item, too. What differentiates them is their ability (or inability) to communicate functionally independent meaning. Meaning is not included in the definition of a word, thus it can be an independent or a dependent lexical item. If it has as an independent meaning of its own, it is a unit of meaning. Since a single word can be a unit of meaning, there is a need for another term in cases where more complex units of meaning consisting of more than word are concerned. The term *extended unit of meaning* seems to serve this purpose well.[9]

2.4. Collocation and meaning-shift: from Firth to Sinclair

In his later work (see for example Cheng et al. 2009), Sinclair starts to use a new term in place of a unit of meaning: *a meaning-shift unit*. This new term is in a way advantageous as it is explicit in conveying the key postulate of Sinclair's conceptualisation of lexis and meaning: when several words start to co-occur and become co-selected on the idiom principle, they undergo a meaning-shift, thus, the meaning of the resulting meaning-shift unit (MSU) should not be traced back to the individual meanings of the words comprising it, their 'core' meanings. The core meaning of a word is the meaning which "first comes to mind for most people" when the word is presented alone, it is hypothesised to be "the most frequent

[8] This difference between the two kinds of lexical items is hypothesised to be psycholinguistically relevant too. The hypothesis will be examined in Chapter 6 as part of a more general hypothesis about the psycholinguistic reality of the model of a unit of meaning.

[9] It is hard to say whether Sinclair himself used the term *extended unit of meaning* in exactly this sense, but it seems that the definition given should not be in serious conflict with his conception.

independent sense" of a word (Sinclair 1987:323). However, when this word participates in a unit of meaning, its core meaning delexicalises.[10] In proposing this meaning-shift, Sinclair, who is considered to be a follower of Firth and the major representative of 'neo-Firthian' tradition in corpus linguistics, departs from Firth's thinking. To explain the concept of a meaning-shift, I will go back to Firth and try to show the important difference of Sinclair's approach to meaning and collocation.

The term collocation is usually ascribed to Firth who worked before the advent of corpus linguistics. For Firth, collocation is a mode of meaning, along with the phonetic, phonological, prosodic and grammatical modes: it is a way "to make statements of meaning" (Firth 1957 [1951]: 192). "Meaning by collocation" is also a way of avoiding ostensive definition or "a language of 'shifted terms'" (Firth 1968 [1957]: 177) which Firth strongly opposes "since the main purpose is the exposition of linguistics as a discipline and technique for the statement of meanings without reference to such dualisms and dichotomies as word and idea, overt expressions and covert concepts, language and thought, subject and object" (Firth 1957 [1951]: 192).

Importantly, for Firth the main function of collocation as well as all the other modes is disambiguation of meaning: it helps to interpret the meaning of a word and distinguishes it from other (similar) words. One of his most famous examples of a collocation is the following:

> It can safely be stated that part of the 'meaning' of *cows* can be indicated by such collocations as *They are milking the cows*, *Cows give milk*. The words *tigresses* or *lionesses* are not so collocated and are already clearly separated in meaning at the *collocational level*. (Firth 1968 [1957]: 180)

Citing this quotation, Geeraerts writes: "This observation is taken as a methodological starting point [in distributional corpus analysis]: the words co-occurring with another one help to identify the properties of the word under scrutiny" (Geeraerts 2010: 169). This seems to be true for most of corpus linguistic studies investigating lexical patterns. Firth's famous dictum "You shall know a word by the company it keeps" seems to be the motto of the field (Firth 1968 [1957]: 179).

However, let us have a closer look at the example Firth provides to illustrate the dictum, which is in fact cited much less often. In the example, Firth examines the meaning of

[10] The process of delexicalisation also obscures the association between a word with its core meaning and the same word as a component of a larger unit (see Section 2.6.4 for the discussion of the relationship between semantic prosody and intuition).

the word ass, which in his view should be read from its immediate context: since “a text in [such] established usage may contain sentences such as ‘Don’t be such an ass!’, ‘You silly ass!’, ‘What an ass he is!’ [...] one of the meanings of ass is its habitual collocation with such other words as those above quoted” (Firth 1968 [1957]: 179). That is, he takes the word as a starting point for the search of meanings it can express in different co-texts. Yet, equipped with corpus linguistic tools, we can approach the task from a different direction and instead of taking form for granted and pair it with all the meanings it can express, we can first try to identify the most invariable meaning and then pair it with the form which consistently expresses this meaning. The word *ass* itself does not mean ‘you are being foolish’, it means “a hoofed mammal of the horse family" (*Oxford Dictionary of English* 2010). It is the co-occurrence of *ass* with a human referent which evokes the meaning of ‘foolish’. Since form can be variable and yet have largely the same meaning, instead of trying to find the most invariable form and map it on all the possible meanings it can express, it might be more useful to look for the most invariable, functionally independent meaning.

In contrast, for Firth “[t]he habitual collocations in which words under study appear are quite simply the mere word accompaniment, the other word-material in which they are most commonly or most characteristically embedded” (Firth 1968 [1957]: 180). However, even in his time Firth predicts: “[i]t will then be found that meaning by collocation will suggest a small number of groups of collocations for each word studied. The next step is the choice of definitions for meanings suggested by the groups” (Firth 1968 [1957]: 181). Sinclair takes that step and states that meaning arises from such “groups of collocations” because collocations in a group are co-selected, and a group, which he models as a unit of meaning, has an independent meaning of its own, which may have nothing to do with the meanings of the words comprising it.

In an interview with Wolfgang Teubert prefacing the publication of the OSTI report[11] (Sinclair et al. 2004) originally written in 1970, Sinclair explicitly draws a difference between his idea of collocation and that of Firth. While for Firth “[o]ne of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation with *night*” (Firth 1957: 196), for Sinclair, as he states himself, “[t]he phrase *dark night* has its own meaning” (Sinclair et al. 2004: xxi). In Cheng, Greaves, Sinclair and Warren 2009 this point is made even clearer:

> …when writers and speakers co-select words, they create a new meaning which makes other instances of the same individual words and other co-selections involving

[11] Originally a Report of The University of Birmingham to the UK Government Office of Scientific and Technical Information (OSTI), entitled “English Lexical Studies”.

> these same words irrelevant. Accordingly, when a co-occurrence, such as ‘hard + work’, is deemed to be significant, the instances of co-occurrence of ‘hard’ and ‘work’ ‘are no longer separate or separable linguistic entities, and their behaviour is entirely accounted for in their membership of the new unit’. All the other instances of ‘hard’ and ‘work’ in the text or corpus ‘are completely irrelevant, being merely homographs’ (Sinclair 2007: 1), because these other occurrences are not co-selected in the unit of meaning ‘hard + work’ and they are in fact members of other units of meaning, each of which is comprised of a unique co-selection of words. (Cheng et al. 2009: 237)

Whenever *dark* and *night* (or *hard* and *work*) are co-selected, these are not two separate words any more but a phrase which has a meaning of its own. That is, *dark* and *night* separately are not relevant for the analysis *dark night* as a unit because both of the words have undergone a certain delexicalisation when they become co-selected as a phrase. Delexicalisation of words participating in a unit of meaning is a matter of degree: while it is not that obvious that a meaning shift has occurred in a unit like *dark night* or *hard work*, it is commonly acknowledged in a phrase like *on the one/other hand*: *hand* as a part of a body is not evoked in the phrase, although if we stop and think, we can of course track the phrase back to the ‘original’ meaning of the word *hand*, which participates in the phrase. In this way, we can posit a continuum of delexicalisation instead of a more traditional continuum between free word combinations and fixed expressions. This proposed continuum is similar to the traditional one in that units of meaning moving along the continuum become more and more fixed as the words comprising them become more and more delexicalised, since the more they are delexicalised, the larger the meaning-shift. However, it is different from the traditional continuum in that it is applicable only to units of meaning produced on the idiom principle: words comprising free word combinations, the ones which are produced on the open-choice principle, are not delexicalised in any way and are outside the continuum.

An important question arises from this conceptualisation: Why would a string of words like *hard work* suddenly start to mean something different from what the sum of the individual words comprising it would normally mean? Something drastic and crucial for the interpretation of meaning has to happen. While delexicalisation is a process, something has to switch it on. It is the idiom principle which occasions the switch.

2.5. Co-selection or the idiom principle

The idiom principle is a key element of Sinclair's conceptualisation of lexis and meaning. It is the idiom principle which causes meaning-shift and the emergence of a new unit, distinct from a word. Co-selection and delexicalisation are two sides of one process: what is co-selected is also delexicalised; delexicalisation then also leads to a meaning-shift.[12]

A combination of linguistic elements becomes a phrase by virtue of being produced on the idiom principle. The fact that all the components of a unit are produced as a result of a single choice ensures that it brings forth just one distinguishable meaning since "meaning arises from choice". In other words, we can count the meanings expressed by the number of choices made and where there is no choice, there is no new meaning (see Sinclair 2004; Sinclair 2008; Sinclair et al. 2004):[13] that is, lexical items produced by syntagmatic association rather than independent choice only form parts of larger units of meaning rather than communicate meanings themselves.

Thus, we can define a unit of meaning as a sequence of lexical items which is produced not as a result of successive paradigmatic lexical choices and application of the rules of grammar but as a single choice of meaning and an activation of internal syntagmatic associations between its elements which glue them together in this sequence – an "occasion where one decision leads to more than one word in text" (Sinclair 1987: 321). To put it more concisely: a unit of meaning is an independent lexical item produced on the idiom principle.

The present account of a unit of meaning includes Sinclair's concept of semantic prosody as one of the key components. The next section will provide an interpretation and an elaboration of the concept.

2.6. Semantic prosody as a communicative function of a unit of meaning

Through the process of meaning-shift, individual components of a unit of meaning become assimilated to each other and acquire a new holistic meaning, semantic prosody. That is, when a lexical item is put to use, it starts to have a meaning which is quite different from the

[12] Basically, grammaticalisation is one kind of a meaning-shift.

[13] "Is there any point in analysing 'Don't count your chickens before they're hatched'? On the one hand, it looks like a well-formed sentence of two clauses, but on the other hand there seem to be hardly any alternatives to the succession of word choices, and a grammatical analysis in such circumstances has little value if one believes that meaning arises from choice" (Sinclair 2004: 132).

"Grammatical meaning is created by choice, and where there is no choice, there is no meaning" (Sinclair 2008: 408).

"If you take information theory, which was, by the way, also anticipated by J. R. Firth, then it tells you that if there is no choice, then there is no meaning" (Sinclair et al. 2004: xxv-xxvi).

dictionary meanings of its component words. Since the item is used for a purpose, its meaning becomes functional and "on the pragmatic side of the semantics/pragmatics continuum" (Sinclair 2004 [1996a]: 34). This kind of meaning is the semantic prosody of a unit of meaning. Therefore, it could be said that, for example, the semantic prosody of the word *presumably* is "a qualified acceptance of the proposition" (Mittwoch, Huddleston and Collins 2002: 769) or hedging. In other words, semantic prosody "expresses something close to the 'function' of the item – it shows how the rest of the item is to be interpreted functionally. Without it, the string just 'means' – it is not put to use in a viable communication" (Sinclair 2004 [1996a]: 34).

The concept of semantic prosody has already been touched upon by now: as an obligatory component of a unit of meaning, as a single choice which determines the co-selection in the idiom principle, as a meaning which characterises a MSU in contrast to a meaning of a word. However, it seems that as it is a rather complex concept, it needs a more detailed treatment of its features. All the more so as it has been a point of continuous debate in the research literature right from the time when it was coined by Sinclair and first introduced by Bill Louw in 1993. Since then, our understanding of the concept and its application seems to have bifurcated in two different directions: one, following Louw's interpretation and the other, trying to track down what Sinclair originally meant by the term. Some scholars propose that we have come to the point where it is important to accept that what we now have are two distinct concepts, and each should have its own place and name in linguistic theory (e.g. Hunston 2007; Stewart 2010). However, virtually all the scholars who comment on semantic prosody refer to both Sinclair and Louw and use all the accumulating research on semantic prosody as a single monolithic whole, which may be misleading.

There are a number of specific questions which different scholars continuously take up in their accounts of semantic prosody. Some of them are: Where does it reside and where does it extend to? How is it different or similar to connotation and semantic preference? Is it evaluative in the first place? Is it a synchronic or a diachronic phenomenon? And lastly: Why is it not available to intuition and introspection? Thus, in what follows, the most important aspects of the concept will be discussed in the light of these specific questions.

The starting point for the discussion will be quite uncompromising: in this study, semantic prosody is treated as an integral element of Sinclair's conceptualisation of lexis and meaning, which is looked at as one system. The system works only if all of its elements are taken into account and therefore, semantic prosody is inseparable from the search for units of meaning. Semantic prosody developed by Bill Louw is a different concept altogether and should be clearly separated from Sinclair's idea (an interpretation of the concept suggested by Louw will be presented in Section 2.9). Then it becomes clear that only an independent unit of meaning can be characterised by semantic prosody.

2.6.1. Semantic prosody. Where does it belong?

The first question we will deal with concerns the types of linguistic items that can be characterised by semantic prosody. I am switching to the term 'linguistic item' here instead of the more usual lexical item because some of the items suggested in the literature as bearers of semantic prosody do not occur in natural language use as such, but are generalisations created by linguists, such as, for example a lemma. The question does not always receive explicit commentary in different conceptualisations of semantic prosody, but is arguably of utmost importance since attributing semantic prosody to different types of linguistic items (a word, a lemma, a unit of meaning) seems to be one of the reasons why different authors do not agree on the nature of the phenomenon. Sometimes the authors spell out themselves what kind of linguistic items they treat as bearers of semantic prosody, but sometimes it only becomes apparent from the analysis they carry out. I will start with comparing Louw's and Sinclair's early analyses of semantic prosody and then take a couple of examples from other commentators: Partington; Morley and Partington; Hunston; and Stubbs.

In his 1993 article, where the concept of semantic prosody was mentioned for the first time, Bill Louw gives several examples of cases where semantic prosodies are used for achieving a rhetorical effect. He shows how, for example, by using the word *utterly*, whose right-collocates show "an overwhelmingly 'bad' prosody" in a concordance, Philip Larkin creates "sinister implications" in a poem of his (Louw 1993: 160). In Louw's view this is "a phenomenon similar to that identified for *set in*" since the subjects which occur with this phrasal verb are usually negative, e.g. *rot* or *decay* (Louw 1993: 160).

However, if we go back to Sinclair's analysis of SET IN, we will see that he was not so concerned with finding the prosody of this phrasal verb as with (1) showing that the meaning usually expressed in conjunction with its occurrence is larger than the one enclosed in the verb itself and (2) establishing a new form-meaning pairing which would better account for the corpus data. With these goals in mind, Sinclair (1991) thoroughly analyses *all* the occurrences of SET IN in his data, takes pains to eliminate all the intervening occurrences to be able to concentrate on the usage of SET IN as a phrasal verb only. Further, looking closely at the usage patterns SET IN as a phrasal verb forms, he points out that first, "it seems to occur typically in a small and/or minor part of the sentence", second, the majority of verbal groups are in the narrative past tense or present tense, and third, nine out of ten present tense occurrences of *sets* "deal with general states of affairs rather than here-and-now" (Sinclair 1991:74). Then, he turns to talk about "the nature of the subjects" of the phrasal verb (Sinclair 1991:74). In addition to pointing out that these subjects (*rot, decay, malaise, despair, ill-will* etc.) "refer to unpleasant state of affairs", he mentions that they are "largely abstractions: several are nominalizations of another part of speech" (Sinclair 1991:75).

In other words, Sinclair's analysis of SET IN is an exercise of pairing a recurring form, however variable and abstracted it can be, with the most consistent, invariable meaning, which is, importantly, not confined to negativity or positivity. This way he establishes an independent unit of meaning which turns out to be much larger than the phrasal verb itself.

What he shows is that SET IN used as a phrasal verb is only a part of a larger, extended unit of meaning whose existence can be demonstrated by the regularity of its pattern. And what is more, this larger, extended unit of meaning has a larger meaning of its own which is expressed whenever the unit occurs in its attested pattern: "If something unpleasant *sets in*, it begins and seems likely to continue and develop", citing the entry for the item given in the *Collins COBUILD English Language Dictionary* (Sinclair 1991:75). This larger meaning, which determines all the elements of the pattern including the negativity of the subjects of the core, can be suggested as the semantic prosody of an extended unit of meaning with the core SET IN, even though Sinclair does not yet use the term himself. In the concluding remarks, Sinclair stresses that "[i]nstead of individual words and phrases being crudely associated with a 'meaning', we could see them presented in active and typical contexts" (Sinclair 1991:78).

The problem of identifying an item which can be associated with a meaning of its own is often the stumbling block. For example, Hunston (2007) writes that both Sinclair and Partington "take as their starting point the individual word (e.g. *budge* or *brook* for Sinclair, *happen* or *sheer* for Partington), and both stress the fact that meaning belongs to a unit that is larger than the word" (250). However, it could be argued that, in talking about the semantic prosody of BUDGE, Sinclair is just taking a shortcut since BUDGE takes part in only one distinct unit of meaning. At the same time, it is problematic to talk about the semantic prosody of HAPPEN, since HAPPEN takes part in a number of different units of meaning and therefore cannot be referred to as a unit of meaning. Perhaps this is why it seems that Partington, as Hunston puts it, "prioritises semantic prosody as the property of a word, and as a feature that distinguishes near-synonyms, whereas Sinclair stresses that the word is only the core of a longer sequence of co-occurring items comprising a 'unit of meaning'" (Hunston 2007: 250). This difference in conceptualisation seems to remain in Alan Partington's 2009 article co-authored with John Morley where the concept of semantic prosody is used to show why the writer preferred the word *peddled* to the word *advocated* and is thus argued to be valuable "in distinguishing among items considered to be synonyms or translation equivalents" (Morley and Partington 2009:140).

This, I think, is a general problem: often researchers concentrate on a lemma of a frequent verb and for some reason assume that it will always be the core of one and the same unit of meaning. Therefore, the chance of running into the problem of counter-examples is highly likely. But if we agree that a lemma is not a lexical unit and that one form correlates with just one meaning, there will be no counter-examples. For example, Hunston (2007) thoroughly analyses a well-known example of CAUSE (first given by Stubbs 1995) which is said to have an unfavourable semantic prosody because the things which are caused are almost exclusively negative, such as *damage, problems, misery*. However, Hunston points out that the use of CAUSE in academic genre is neutral and proposes that "CAUSE implies something undesirable only when human beings, or at least animate beings, are clearly involved" (253). Stubbs (2009) seemingly agrees with Hunston but by saying that "in scientific and technical texts the semantic preference and the semantic prosody are likely to be cancelled" (130). However, there could be a different explanation of this counter-example: whenever CAUSE is used in academic context and does not deal with human beings, i.e. acquires a different semantic preference, it just enters into a different unit of meaning which naturally has a different meaning, a different semantic prosody. A change in form, in this case in the semantic preference, leads to a change in meaning.

As Hunston points out, "ascribing semantic prosody to a word is over-simplistic" since:

> If the phraseology changes, the semantic prosody is also different. This is not particularly surprising, but it serves as a useful reminder that, in Sinclair's examples at least, *semantic prosody is a discourse function of a sequence* rather than a property of a word. (Hunston 2007: 258; emphasis mine)

In other words, for Sinclair each different use brings forth a different meaning as well, so for example *take place* has nothing to do with *take*, or no more than *take* has to do with *teach*, although they both start with the letter *t* (see Sinclair 1991: 78 the discussion of *set* and *set in train*). So if we would like to follow Sinclair, we would first need to identify a unit of meaning and only then talk about its semantic prosody.

2.6.2. Semantic prosody, connotation and evaluation

There is a widespread view that semantic prosody is first and foremost evaluative. This view seems to either draw on Sinclair's words that "[a] semantic prosody is attitudinal, and on the pragmatic side of the semantics/pragmatics continuum" (Sinclair 2004 [1996a]: 34) or that "[i]t is a subtle element of attitudinal, often pragmatic meaning..." (Sinclair 2004 [1998]: 144) or build on Louw's concept of semantic prosody which is characterised by either positive or negative speaker/writer attitude (see e.g. Louw 2000).[14] In particular, Morley and Partington (2009) describe semantic prosody as essentially evaluative and argue that "[s]emantic prosody is an expression of the innate human need and desire to evaluate entities in the world they inhabit as essentially *good* or *bad*" and that "evaluation at its most basic is a two-term system" (141). Interestingly, bad prosodies have been detected much more often than good ones, which is suggestive of the idea sometimes expressed that it is the negative state of affairs which drives the need for communication (Louw 2000; Partington 2004: 133; see also Stewart 2010: 46).

Indeed, if we agree with Sinclair that semantic prosody is an *obligatory* component of a unit of meaning and at the same time argue that it is inherently evaluative, we are bound to see everything as evaluative,[15] which inevitably leads us to the kind of conclusions described above. If we reject such vision, but still argue that semantic prosodies are evaluative, we must accept that not all units of meaning have semantic prosodies. So, is semantic prosody evaluative? What if evaluation is just one type of the functions semantic prosody can realise?

For Partington (1998), for example, semantic prosody is a kind of expressive connotation (along with social and cultural) which is "not contained in a single item, but is expressed by that item in association with others, with its collocates" (66). In other words, in Partington's account semantic prosody is not given "the leading role to play" but is subsumed in the category of connotation. I would suggest that the hierarchy is reversed: as the initial reason for choosing a unit of meaning, whether it is a single word or a multi-word unit, semantic prosody seems to be functioning at a higher level than connotation. All the meanings a unit of meaning has (denotational, connotational) contribute to it being chosen as suitable to fulfil the required semantic prosody. As a communicative function of a unit of meaning, semantic prosody seems to be close to the concept of illocutionary force.

[14] But as I will be arguing, Bill Louw's semantic prosody is a different concept altogether and can be termed something like 'logical prosody', yet the terminological confusion and lumping everything that has ever been said about semantic prosody into one theory very often leads to simplification in understanding Sinclair's or Louw's semantic prosody and in ultimate cases to a reduction ad absurdum.

[15] See Mauranen 2004 for the discussion of the problems a broad definition of evaluation as 'omnipresent' creates.

As Partington (1998) points out, there are words whose expressive connotation is the most substantial part of their meaning:[16] "A word like *pig-headed* only exists because it has an expressive connotation of disapproval ... Similarly, the sole purpose of the term *venerable* is to put old age in a good light, and that of *callow* to express disapproval of youth" (66). In such cases, I would completely agree, semantic prosody is evaluative: these words come to be chosen because their expressive connotation is the reason for their employment and therefore aligns with their semantic prosody. However, it does not follow that semantic prosody as a concept equals expressive connotation since not all units of meaning are inherently evaluative. Rather, the value of semantic prosody is closest to the value of expressive connotation when the purpose of the message is attitudinal or evaluative and a single word constitutes a unit of meaning.

The case of *utterly* can perhaps illustrate the point. As an intensifier it co-occurs with adjectives and adverbs. It has been noted (e.g. Louw 1993) that the nature of these co-occurring adjectives and adverbs is predominately negative. Without going into details, we may postulate a unit of meaning with the core *utterly*[17] and a colligation with adjectives and adverbs. The semantic prosody of this unit of meaning, which is comprised of *utterly* and an adjective it modifies, would be something like 'to express a negative attitude towards the high degree with which some quality shows itself in something'. In other words, there is no denying that the semantic prosody of this unit of meaning is evaluative but with a different underlying logic: not because all semantic prosodies are evaluative but because negative evaluation is intrinsic to *this* unit in particular and hence, the purpose for its use cannot be anything else but expressing negative attitude towards something. Again, it is important to keep apart evaluation and semantic prosody even where evaluation is the chief function and meaning of an item: semantic prosody is the functional meaning, and evaluation can be a function in a particular case.

Evaluative prosody (Morley and Partington 2009), which is then a different phenomenon and perhaps the one to be recognised along with semantic prosody, is by

[16] In a similar vein, for example Hunston and Thompson (2000) also mention that "some lexical items are very clearly evaluative, in the sense that evaluation is their chief function and meaning" (14). As examples they give adjectives such as *splendid, terrible, surprising*, adverbs *happily, unfortunately, plainly*, nouns *success, failure, triumph*, verbs *win, loose, doubt.*

[17] Basically, it is counter-intuitive that an intensifier is the core and the adjective it modifies is a colligation, i.e. a co-occurring component, yet the requirement for the core is that it is the most invariable formal part of a unit, therefore a negative adjective is not suitable for the purpose. At the same time, alleging that *utterly* is the core of a unit does not necessarily mean that it is chosen first and the co-occurring adjective is only chosen by association.

definition not obligatory, hence it is closer to the optional categories like collocation and semantic preference. Hunston suggests “attitudinal preference” for this phenomenon:

> the term ‘semantic prosody’ is best restricted to Sinclair’s use of it to refer to the discourse function of a unit of meaning, something that is resistant to precise articulation and that may well not be definable as simply ‘positive’ or ‘negative’. I would suggest that a different term, such as ‘semantic preference’ or perhaps ‘attitudinal preference’, is used to refer to the frequent co-occurrence of a lexical item with items expressing a particular evaluative meaning. (Hunston 2007: 266)

So in our example of a unit with the core *utterly*, the fact that its co-occurring adjectives or adverbs are usually negative would be a realisation of the attitudinal preference of the unit. Adoption of a new category, attitudinal preference, as an association of the core with negative or positive items would also help us to draw a clear-cut distinction between semantic prosody and semantic preference. Semantic preference is an optional component of a unit of meaning: in some units it will be realised, in others not. In the same way, attitudinal preference does not have to be present in every unit of meaning and we do not have to stretch our notion of evaluation to seeing everything as negatively or positively charged. In contrast, semantic prosody is obligatory: a unit of meaning cannot exist without semantic prosody like a form-meaning pairing cannot exist without meaning.

2.6.3. Semantic prosody: Synchronic vs. diachronic perspective

Another frequent topic of studies on semantic prosody is the question whether semantic prosody should be discussed from a synchronic or a diachronic perspective. It is taken up by, for example, Whitsitt,[18] who understood semantic prosody as a flow of meaning “from strong, full, bad words, into the weak, empty, innocent forms” (2005: 292), and hence, claimed that a phenomenon diachronic in its nature cannot be studied by using synchronic corpora. This view is based on the idea that a word by frequent co-occurrence with words having negative connotations i.e. “by keeping bad company” acquires a negative semantic prosody which is then nothing else but a transfer of negative connotation from one word to the other.

Putting aside the evaluative component of this definition, which was discussed in the previous section, I will concentrate on the synchronic versus diachronic dimension of the concept. Commenting on Louw’s suggestion of the notion of “contagion”, Sinclair writes the following: “...with frequent usage together, words form syntagmatic associations with others

[18] See also Stewart (2010): “a shift of meaning during the course of time”, “semantic change” (43).

round them, so that instead of merely taking on some of the meaning of their surroundings through contagion, *they form a new unit of meaning* which requires the presence of both words (or more than two in many cases) to be instantiated" (Sinclair 2004 [1996a]: 150, emphasis mine). Here two things are crucial: first, it is useful to remember the distinction between Firth's and Sinclair's ideas of collocation discussed in Section 2.4: when *dark* and *night* co-occur, they do not restrict the meaning of each other but form a new unit with a new meaning. The second postulate essential for the argument is that a unit of meaning is a sequence produced on the idiom principle. The idiom principle or co-selection is a psycholinguistic mechanism of language production, and therefore strictly synchronic: we speak synchronically, not diachronically. So the fact that *dark* and *night* were previously separate items which started to co-occur is irrelevant for the production on the idiom principle. We may be able to infer the etymology of a unit, but it is not what makes us use the unit in the first place. The unit has acquired its semantic prosody through meaning-shift, but this meaning-shift or the diachronic perspective on the emergence of a unit of meaning is irrelevant at the time of use.

2.6.4. Semantic prosody and intuition

It is often stated that semantic prosodies are not available to our intuition or conscious knowledge (see Stewart 2010 for a comprehensive overview). Indeed, the mysterious aura created around semantic prosody may be taken to imply that language production as well as understanding is completely subconscious. So what is it exactly that is hidden and how exactly corpus linguistics methodology helps us to reveal it?

Reporting the results of the COBUILD project, Sinclair wrote that one of the problems which arose is that "[t]he commonest meanings of the commonest words are not the meanings supplied by introspection" (Sinclair 1987: 322). As an example he gives the word *back* whose sense 'a part of human body' is not the one frequently used (but, it will be argued, the most independent one). Another example of a less common word is the word *pursue*: "the first sense offered in CED for pursue is 'to follow (a fugitive etc.) in order to capture or overtake', yet by far the commonest meaning is the fifth sense 'to apply oneself to (one's studies, hobbies, interests etc.)'" (Sinclair 1987: 323).

On the face of such evidence, Sinclair makes several tentative conclusions:

(1) Frequent words or frequent senses of words tend "to have less of a clear an independent meaning".

(2) “This dependency of meaning correlates with the operation of the idiom principle to make fewer and larger choices.”

(3) “The ‘core’ meaning of a word – the one that first comes to mind for most people – will not normally be a delexical one. A likely hypothesis is that the ‘core’ meaning is the most frequent independent sense.” (Sinclair 1987:323)

That is, in its most frequent uses a word tends to be co-selected with other elements on which it becomes dependent for the holistic meaning they collectively communicate. Together with losing the independence of meaning, it dissociates with the core meaning it previously had. In lexicographical practices, the meaning of a new unit of meaning a word participates in is usually counted as one of the senses of this word. However, it seems that psycholinguistically this inferred sense does not directly associate with the word. The meaning which does associate is the one which a word can communicate independently, i.e. its core meaning. So what is not available to intuition is not semantic prosody, the meaning of a newly formed unit of meaning or MSU, but the association between a word and the pattern it participates in, its delexical use.

For example, the word *hand* has a meaning as a part of the human body. For a language user, it does not directly associate with *on the one hand* which is not just a different use of *hand* but a different lexical unit altogether. When presented with the word *hand*, we cannot predict that it can be used to compare two opposing factors or views because it cannot; this meaning is expressed by a different unit. Corpus linguistic methodology helps us to see that this different unit exists, as the recurrence of the pattern correlates with a different meaning communicated, i.e. the form-meaning connection is consistent.

Other scholars also mention the unreliability of our intuition in regard to delexicalised uses of words. For example,[19] Stubbs points out that:

> Native speakers have strong and reliable intuitions that some words are more frequent than others: there is not much doubt that *luck* is more frequent than *lute*. But native speakers are unreliable in judging the most frequent uses of frequent words, for example that TAKE is most frequent in its delexicalized uses in phrases such as *take place* and *take a photograph*. (Stubbs 2001: 72)

On the basis of such emerging evidence that our intuitions are unreliable in predicting the usage pattern of a frequent word, we can draw a tentative conclusion that the ‘mental lexicon’

[19] See also Renouf (1987: 174-175) for the frequent uses of KEEP which are not the ‘core’ meanings.

is organised according to meanings rather than according to words e.g. *take* with the core meaning is stored separately from *take place* and again separately from *take a photograph* because these larger units have their own meanings. In other words, these three units are not stored together even though all of them include *take* as a component, and this could be the reason why language users do not arrive at *take a photograph* when presented with the stimulus *take*.

Stubbs also makes some further observations about intuition:

> Native speakers are quite unable to generate, from their intuition alone, comprehensive lists of the most frequent uses of frequent words. However, although native speakers have no intuitive access to such information, when they see automatically generated lists of recurrent n-grams, they immediately recognize idiomatic ways of expressing common pragmatic meanings. As Fox (1987: 146) puts it, as soon as they are told what the most frequent uses are, they cannot understand why they did not think of them in the first place: 'the important thing, of course, is that they had not'. (Stubbs 2007: 171)

So, what is inaccessible to intuition or retrospection is the pattern of use and not the meaning. When the pattern is presented in whole, its meaning is obvious; it cannot be recognized on the basis of just one word that participates in the pattern. The likelihood that the pattern can be predicted depends on the degree of delexicalisation of the words participating in the pattern. Yet, it is of course less of a problem if the unit consists of just one word. In this respect it is worth remembering that the eye-opening examples of semantic prosody have been those which describe extended units of meaning, i.e. units consisting of more than one word, rather than relatively 'independent' single-word units. In this way the comparison of the semantic prosodies of BENT ON versus SNOBBISH in order to show that the latter is more negative by just describing the degree of unpleasantness of immediate collocates (Stewart 2010: 34-38) is unjustified. SNOBBISH as an adjective has a relatively clear meaning while BENT ON participates in a larger unit whose pattern is only revealed through concordancing. It was Louw (1993) who took up the pattern of BENT ON originally. Although he does not analyse the usage pattern of the phrasal verb in terms of a unit of meaning and its components, he does mention that "the pursuits that people are BENT ON are almost always negative or unpleasant in some way" (Louw 1993: 166). That is, it is the *people* that *are* bent on, and it is *their pursuits* that they are bent on, and what is more, it is *the pursuits* that are negative (which are also usually expressed by abstract nouns or *ing*-forms) and not just the left or right collocates. And the reason why the news about the semantic prosody of BENT ON comes more unexpected than that of SNOBBISH is the fact that BENT ON is only the core of a larger unit of meaning whose pattern is not self-evident and can only be observed by arranging the instances of its occurrence vertically, i.e. by concordancing. So, what is hidden, or to be more exact, not prompted by a single-word stimulus, is actually not the semantic prosody but the unit of meaning, the pattern.

In fact, if we accept the idea that language production on the idiom principle is a psycholinguistic mechanism which is characterised by automaticity, i.e. "absence of attentional control in the execution of a cognitive activity" (Segalowitz and Hulstijn 2009: 371) and therefore, a mechanism which makes use of implicit memory or knowledge,[20] it becomes clear that something which is produced unconsciously or automatically may not be retrievable consciously.

To sum up the discussion of Section 2.6, in this thesis semantic prosody is regarded as an obligatory meaning-component of a unit of meaning: it is its communicative function. Therefore, it is considered that (1) semantic prosody is a property of an independent lexical item - a unit of meaning; (2) it is not inherently evaluative or attitudinal, even though sometimes the evaluative component of meaning becomes most prominent; (3) it is relevant at the time of use and therefore analysed from a synchronic perspective; and (4) it is not 'hidden': it is the patterning of a unit of meaning which cannot be predicted on the basis of a single word participating in it, semantic prosody can normally be read when the whole unit of meaning it characterises is presented.

2.7. The theory of meaning and the ultimate dictionary

A unit of meaning with its five components, two obligatory (the core, the semantic prosody) and three optional (collocation, colligation and semantic preference), was conceptualised by Sinclair as a basic unit of the ultimate dictionary. The obligatory components constitute the backbone of the form-meaning pairing, while the optional categories serve to "fine-tune the meaning" and "give semantic cohesion to the text as a whole" (Sinclair 2004 [1998]: 141). In a way, the model aspires to implement the assertion that "form and meaning cannot be separated because they are the same thing" (Sinclair 2004 [1998]: 139).

Sinclair criticised the contemporary model of a dictionary which "has always been based on the rough equation of a word and a unit of meaning" (Sinclair 2004 [1998]: 132).

[20] Cf. Ellis (2002): "To the extent that language processing is based on frequency and probabilistic knowledge, language learning is implicit learning" (145).

He pointed out that there was no theory that would allow for phrases, rather than treat them as a nuisance, and allow "for the relationship between the 'independent' and the 'dependent' uses of a word" (Sinclair 2004 [1998]: 132). Indeed, "the idea that a word could inherently have one or more meanings" (132) proved to be shaky: at the very least the assumed polysemy of words makes it hard to explain how language manages to be unambiguous.[21] Instead the far-outweighing importance of the surrounding language suggests that "many, if not most, meanings require the presence of more than one word for their normal realization"; and that "patterns of co-selection among words [...] have a direct connection with meaning" (Sinclair 2004 [1998]: 132-133).

The ultimate dictionary would then list units of meaning rather than words since, as Sinclair's research shows:

> For every distinct unit of meaning there is a full phrasal expression which is differentiated from all other full expressions of units of meaning, and which we call the canonical form. [...] A dictionary containing all the lexical items of a language, each one in its canonical form with a list of possible variations, would be the ultimate dictionary. (Sinclair et al. 2004: xxiv)

Sinclair also gives an example of such canonical forms that can make an entry in the dictionary: the phrase *get in touch with* would be a kind of a prototype for a unit of meaning with the invariable core *in touch with*, and the default collocate *get*, which can be replaced with other verbs like *bring, be, keep, remain* (Sinclair et al. 2004: xxiv). Cheng et al. (2009) explore the idea of a canonical form of a unit of meaning, or, to be exact a meaning-shift unit, with a phraseological tool ConcGram (Greaves 2009 [2006]). For the two positional variants of an automatically retrieved concgram,[22] a co-occurrence of two words irrespective of positional or constituency variation in a specified span (see Section 4.2.2 for more details), PLAY/ROLE and ROLE/PLAY, they establish two distinct canonical forms. One of them constitutes a contiguous textual object, a unit of meaning whose constituent elements combine to form a single entity through endocentric relationship, '*role* PLAY' which expresses the meaning "some kind of activity in which the participants take on contrived roles as part of some form of training programme or entertainment" (Cheng et al. 2009: 247). The second canonical form PLAY*ROLE[23] comprises a textual incident,[24] a unit of meaning

[21] See also Teubert (2005): "Once we replace the concept of the polysemous single word by the concept of the monosemous lexical item, the problem of ambiguity […] suddenly disappears" (6).

[22] Cheng et al. used a 5-million subset of the BNC.

[23] Here as in Cheng et al. one asterisk stands for one intervening word. In contrast, as a rule I use an asterisk for zero or more characters as in the query syntax of the BNC.

whose constituent elements remain separate to construct a meaning through exocentric relationship. It has a number of configurations (e.g. PLAY***/****ROLE, ROLE*PLAY) but all of them realise one distinct canonical meaning, or the semantic prosody of "a 'weighty/meaningful' thing to do" (Cheng et al. 2009: 245) with different "degrees of turbulence". The formal variation which occurs inside the MSU or unit of meaning is also described in terms of collocation, colligation and semantic preference.

The idea of a canonical form of a unit of meaning is the solution to the problem of how to relate a finite set of meaningful items to the unlimited set of meanings in use (Sinclair 2004 [1998]: 134). Yet, the problem is complicated by the fact that "some aspects of textual meaning arise from the particular combinations of choices at one place in the text, and there is no place in the lexicon-grammar model where such meaning can be assigned", which in essence means that our present lexicons are "doomed" (Sinclair 2004 [1998]: 134). Therefore, "the units of the 'live' lexicon ... adapt to the ever-changing, never-quite-repeated circumstances of communication, and as such cannot, in principle, be fully prescribed in advance" (Sinclair 2004 [1996b]: 161). The "live" lexicon is, hence, "empty" as it has to learn about vocabulary from texts and be constantly updated (Sinclair 2004 [1996b]:162).

While the purpose of all the previous sections has been to explain Sinclair's account of lexis and meaning with all its concepts – a unit of meaning, collocation, colligation, semantic preference, semantic prosody, meaning-shift, canonical form, idiom and open-choice principles – in the form they are adopted in this study, the rest of the sections will juxtapose Sinclair's approach to language patterning with some other closely related approaches. I will start with Hoey's lexical priming, then move on to Louw's semantic prosody, compare Sinclair's approach to phraseology with Wray's notion of formulaicity and conclude by discussing the implications of the argument about the psycholinguistic reality of a unit of meaning pursued in this thesis. It is hoped that these comparative discussions will be able to further clarify some of the aspects of Sinclair's conceptualisation.

2.8. Lexical priming

Michael Hoey takes a psycholinguistic perspective on co-occurrence phenomena in his lexical priming theory. The main idea of the theory is that "the explanation of the phenomena

[24] The distinction between *textual objects* and *textual incidents* is proposed in Sinclair and Mauranen (2006: 149, 154-155). Cheng et al. (2009) employ the concepts for the analysis of concgrams and the canonical forms they shape.

Sinclair describes - collocation, colligation, and semantic preference (which Hoey calls semantic association) - lies in the process of acquisition" (Hoey and O'Donnell 2008: 295).

Hoey's main argument is that it is a human being who operates on the idiom principle, not the language itself: "words are never primed *per se*; they are only primed for someone" (Hoey 2005: 15, emphasis in the original). By lexical priming Hoey means that "whenever we encounter a word, syllable or combination of words we note subconsciously (1) the words it occurs with (its *collocations*), the meanings with which it is associated (its *semantic associations*), the grammatical patterns it is associated with (its *colligations*), and the interactive functions it contributes to serving (its *pragmatic associations*)", (2) "the genre and/or style and/or social situation it is used in", (3) "its text-linguistic characteristics: the positions in a text that it occurs in (its textual colligations), the cohesion it favours or avoids (its textual collocations) and the textual relations it contributes to forming (its textual semantic associations)" (Hoey 2009: 34-35). And "when we come to use the word (or syllable or cluster) ourselves" (Hoey 2009: 36), we tend to reproduce all the contexts and co-texts it has become "cumulatively loaded with" (Hoey 2005: 8). In other words, our lexical primings or the tendency to use lexical items in particular contexts and co-texts emerge as a result of our own individual experience with the language in the specific domains we actually use it. This experience is stored in the form of "mental concordance":

> …the mind has a mental concordance of every word it has encountered, a concordance that has been richly glossed for social, physical, discoursal, generic and interpersonal context. The mental concordance is accessible and can be processed in much the same way that a computer concordance is, so that all kinds of patterns, including collocational patterns, are available for use. (Hoey 2005: 11)

Lexical priming theory seems reasonable in many ways: linguistic experience (or "input") plays a major role in language acquisition and performance. It is also acceptable to common sense that we tend to memorise and come to associate those things that happen together. Since we experience language in a linear fashion, it is likely that we come to associate items which follow each other. We can also become biased towards the patterns we are familiar with. However, there are some problems in the model which might question its psycholinguistic validity.

2.8.1. The importance of meaning for the psycholinguistic reality

At the root of the argument, Hoey states that in the theory of lexical priming he is developing Sinclair's ideas on the pervasiveness of co-occurrence patterns in text, especially their psycholinguistic dimension. However, as argued in the previous sections, the elements of Sinclair's conceptualisation are inextricably intertwined with one presupposing another. Hoey, in contrast, adopts the categories of co-occurrence, such as colligation and semantic preference, but seemingly leaves aside the concept of a unit of meaning as the locus of these co-occurrences brought about by the single unifying meaning.

In particular, Hoey defines collocation as "a psychological association between words (rather than lemmas) up to four words apart […] evidenced by their occurrence together in corpora more often than is explicable in terms of random distribution" (Hoey 2005: 5). However, if the only criterion for the priming to be established between two words (or a word and a grammatical class, or a word and a semantic set etc.) is the recurrent co-occurrence, then we can get quite odd pairings which psycholinguistically do not seem to be plausible. For example, the most frequent bigrams in the BNC calculated by "Phrases in English" (Fletcher) are *of the, in the, to the, on the*. Would that mean that we are primed to associate *the* with *of*, *in*, *to* and *on* with different intensity, or would that mean that we are primed to colligate *the* with the class of prepositions? The point is: until meaning is taken into account, lexical patterning claims do not appear to be psycholinguistically relevant. It is the fact that words share one meaning that seems to keep them together in the mind. Human pattern-finding is strongly connected to intention-reading (Tomasello 2003). Meaningless co-occurrences must have a much smaller chance of acquiring representation in the mind. This does not mean that syntagmatic association is non-existent; however, the argument goes, hypothetically, it mostly operates *inside* a meaningful unit. The hypothesis will be studied in Chapter 6, which discusses the psycholinguistic reality of a unit of meaning.

2.8.2. Dependent choices at different levels: psycholinguistic vs. other

My second argument is closely related to the first, i.e. the importance of identifying a unit of meaning. Hoey seems to start from the assumption that all co-occurrences which are observable with corpus linguistic methods and are statistically significant have psycholinguistic reality, in other words it is predicted that everything that recurrently co-occurs has a direct effect for storage. However, it seems reasonable to suppose that not only storage mechanisms are able to condition linguistic choices and create predictability but there are also, for example, social and logical constrains which are not necessarily stored. For example, an apparent existence of textual associations may be easier to explain if we accept the possibility of conditioning at more abstract levels. At the same time there is no denying that such more abstract associations can also be stored for someone: the essential point that priming is individual predicts that all language users (of a particular speech community) do not necessarily share the same primings, yet for some, certain abstract choices can also be automatised. Nevertheless, hypothetically, lexical choices are more likely candidates for systematic storage.

In other words, it seems useful to be prepared to admit that everything which shows a pattern in corpora, i.e. everything which recurs, does not have to be psycholinguistically real in order to be valid, and real, otherwise. A very good example of such patterns are *semantic sequences* proposed by Susan Hunston (2008) as "another candidate for the description of regularity in text" (Hunston 2010 [2008]:7). She defines them as "recurring sequences of words and phrases that may be very diverse in form and which are therefore more usually characterized as sequences of meaning elements rather than as formal sequences" (Hunston 2010 [2008]:7). For example, Hunston shows that the structure *the* + NOUN + *that, e.g. the idea that, the discovery that, the observation that*, is consistently used for evaluation of "the epistemic status of the proposition expressed in the *that*-clause" in scientific discourse (Hunston 2010 [2008]: 14). Such co-occurrence seems to be related to how we logically formulate our arguments rather than to the storage of the structure *the* + NOUN + *that*. Comparing semantic sequences with Sinclair's units of meaning, Hunston points out that "[t]hey represent what is often said, not how a word is typically used" (Hunston 2010 [2008]: 27).

Indeed, it seems plausible that a unit of meaning is not the only level where choices are not completely autonomous. There could be a conceptual level where "meaning elements", as Hunston calls them, would form a pattern because this is how we structure reality. Predictability at a textual level is also possible as, for example, genre determines a lot of choices which can appear as co-selected but in fact emerge as a result of following the conventions of a specific genre: research into moves analysis of academic writing can serve as an example (Swales 1990; Mauranen 1993). However, choices at conceptual or textual levels, although predetermined to a greater or lesser degree and therefore predictable, do not have to be lexically associated and psycholinguistically represented in order to be reproduced. In other words, it is not the idiom principle, the mechanism of automatic language production/comprehension enabled by the contents of the implicit memory, which lies behind their co-occurrence, but conceptual, logical, textual, social and other constraints which limit the amount of choices available, making certain features of the text predictable. To put it another way, the patterning of the unit of meaning is predictable because (and when) it is produced on the idiom principle while the pattering of a semantic sequence, for example, is predictable because it is part of our "common epistemological practices", like the practice of construing *discovery* "as being the cause, either of an emotion or of an idea" (Hunston 2011: 92, 97).

In contrast, according to Hoey, lexis itself can be for example "textually primed": "many of the features of a text – its organization, its cohesion, its chunking – are latent in the lexical items we select" (Hoey 2005: 115). As such, he argues, that it is the vocabulary of a sentence from a travel book written by Bill Bryson *In winter Hammerfest is a thirty-hour ride by bus from Oslo...* which primes us to expect that the text will be about Hammerfest rather than winter or Oslo: "[y]our experience of the word sequence *in winter* has led you not [...] to anticipate that a text beginning with this word sequence will be about winter" (Hoey 2005: 115).

Another example of a more abstract association in the theory of lexical priming is a semantic association. It is the author's conformity to his collocational primings, Hoey argues, which ensures the naturalness of the sentence quoted above:

> The collocations just listed interlock. So *hour* collocates with *thirty* but it also collocates with *ride*. Likewise *ride*, in addition to collocating with *hour*, collocates with *by* and *bus*. *Bus* also collocates with *by*. Both *ride* and *bus* collocate with *from*. (Hoey 2005: 6)

At the same time, he continues, a certain linguistic creativity is brought by semantic association. "Primings move out from collocations to semantic associations" (Hoey 2005: 18), and we learn to associate not only *hour* with *thirty* but more generally TIME with NUMBER. Consequently, Bill Bryson's sentence also conforms to the following semantic association:

> SMALL PLACE is a NUMBER-TIME-JOURNEY – (by VEHICLE) – from LARGER PLACE. (Hoey 2005: 18)

In support of the hypothesis, Hoey provides examples from corpora of sentences constructed on the same principle. However, the co-occurrence of the elements in this structure may have a different explanation. Distance is usually measured in terms of time and speed, for which it is necessary to know the type of vehicle, and there are not very many alternatives to express this relationship. Therefore, the structure can eventually end up being expressed in a predictable way without being necessarily represented in the mind as lexically associated.

What is indeed not explicable by such other constraints is the structure and properties of such units as *in winter* and *by bus*. While it is not straightforward whether *in winter* and *by bus* should be analysed as collocations or colligations (though see the argument in favour of the first option in Section 2.2), they clearly form units of meaning. A cross-linguistic analysis may be of interest here. In agglutinative languages Finnish and Russian, the equivalent expressions of *in winter* (*talvella, зимой*) and *by bus* (*bussilla, на автобусе*) are represented by an association of a noun with a specific case (plus a preposition in the case of *by bus* in Russian). Interestingly, while in the equivalents of *by bus* the two languages apply a similar case, in the equivalents of *in winter*, the cases used are quite different, which is of course not surprising as the case systems of the two languages are so different that it is hard to compare them. To be exact, a direct translation of both *bussilla* and *на автобусе* would be '*on a bus'*. The point is that it would be quite implausible to suggest that the respective prepositions in the case of English, and case markings in the case of Russian and Finnish would be selected on an open-choice principle or on the basis of, say, logic. The preference for a particular preposition or a case in the three languages is inexplicable in purely grammatical terms. The processing load with open-choice selections would also be unreasonably high. From the psycholinguistic point of view, "[f]requency of occurrence may lead to independent representation of even so-called regular constructional patterns" (Ellis 2002: 168). Also, there is evidence that Finnish speakers indeed store high-frequency full-form representations separately (Lehtonen and Laine 2003).

Furthermore, on the basis of what has been said but also as an additional point, it is argued that it is more useful to analyse such co-occurrences as units of meaning rather than as two words which are primed to co-occur: *in winter* is not a modification of the word *winter* but is a separate, independent unit of meaning, likewise *by* does not either add or subtract any meaning from the word *bus*, *by bus* is a different unit altogether. *By* and *in* are as dependent as case markings in other languages.

In view of the arguments presented above, I will not adopt the theory of lexical priming as part of the theoretical framework for this study, even though just like Hoey I am interested in the psycholinguistic aspect of lexical patterning observable with corpus linguistic methods. Yet, usage-based theories of language hold that language learning is exemplar-based: "The knowledge underlying fluent, systematic, apparently rule-governed use of language is the learner's entire collection of memories of previously experienced utterances" (Ellis and Larsen-Freeman 2006: 565). Thus, I will use the term *priming* in a wider sense: as the effect of previous language exposure, including one's own language use, on subsequent language use.

2.9. Louw's semantic prosody

In his well-known article "Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies", Bill Louw defines semantic prosody as "a consistent aura of meaning with which a form is imbued by its collocates" (Louw 1993: 157). So first and foremost, it is a collocational phenomenon for him. There are two more arguments which I think are key to understating Louw's conceptualisation, first, that "situational and linguistic contexts are co-extensive" (Louw 2000), and second, that semantic prosodies can serve as a diagnostic tool in uncovering this link between the text and the context of situation by revealing irony or insincerity.

So, how does Louw's notion of semantic prosody differ from the account presented here and how does it relate to it, if it does? To sum up the points of the previous section, it seems that in language processing co-occurrence can be predetermined to a larger or smaller degree at different levels of abstraction. Thus we may eventually be able to postulate a series of dependent choices.[25] While in a unit of meaning we are co-selecting its different components, a semantic sequence emerges because this is how we structure reality and the relations between different concepts. Linguistic choices seem to be able to be conditioned by states of affairs as well.[26] In a nutshell, the act of throwing a ball can only be described by a limited number of ways because it is restricted by the laws of physics, and thus cannot for example fly away into the outer space. We know what is going to happen to the ball, and our linguistic choices are going to be conditioned accordingly. This applies to more abstract 'events' just as well. And, arguably, this is what Bill Louw's semantic prosody shows: certain knowledge that the speaker/writer possesses can influence his/her linguistic choices which in their turn might reveal his/her real attitude/intentions. For instance, one of his examples is the negative prosody of the phrase *symptomatic of* which, when applied to describe the University of Zimbabwe, reveals the true state of affairs in the university in spite of the speaker asserting it has a high reputation (Louw 1993: 43-44).

In a later article Louw (2000) explains how in a poem Hawk Roosting by Ted Hughes linguistic means serve to create an expectation that a disaster is going to happen or in Louw's

[25] According to Ellis (2012a), one of the key features of seeing language as a complex dynamic system is the assumption that "[t]he structures of language emerge from interrelated patterns of experience, social interaction, and cognitive processes". This is able to explain, inter alia, "variation at all levels of linguistic organization" (22).

[26] This is not to argue that reality cannot be construed through linguistic choices. In fact this direction of influence seems to receive much more attention. Yet, reality sets its own limits: there are certain things we cannot say because we cannot even imagine them - we did not talk about Internet before it was invented.

words: "*disaster* is the most likely collocate to the right". *Disaster* here would not just be an item that frequently follows another item because they comprise a unit of meaning or because this is how we usually express relations between events or entities. *Disaster* here is an actual disaster that will happen to a hawk. Basically, from the point of view of dependent choices, the author who is writing the poem knows what is going to happen to a hawk, i.e. he knows how the states of affairs are, and this knowledge influences the linguistic choices he makes, sometimes so subtly that it is enough to create apprehension but not enough to explicitly analyse why that happens: this can only be revealed with the help of corpora.

Consequently, it seems reasonable to suggest that Louw's notion of semantic prosody is entirely different from the interpretation of semantic prosody as a communicative function of a unit of meaning and an integral element of the conceptualisation of meaning adopted here: the two concepts seem to be functioning at different levels of abstraction.

2.10. Formulaicity and novelty vs. idiom and open-choice principles

The concept of *formulaic language* has much in common with the idea of the idiom principle. Wray's definition of a *formulaic sequence* is based on the contrast between a sequence holistically retrieved from memory and one constructed through the application of grammatical rules, much like the distinction between the idiom and open-choice principles:

> a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar. (Wray 2002: 9)

The definition also presupposes psycholinguistic reality of a formulaic sequence: it is a sequence stored holistically in the mind. A similar psycholinguistic representation is suggested for a unit of meaning in this study. However, in spite of these evident similarities, there also seem to be important differences in Wray's and Sinclair's approaches to lexis and text.

While Wray is interested in storage and the lexicon, for Sinclair the lexicon is really "empty" (Sinclair 1996b). Perhaps the distinction is most apparent in the fact that Wray works at identifying formulaic language, that is, distinguishing between formulaic and novel language material. Sinclair, in contrast, prioritises the communicative act itself and seeks to describe text as it unfolds in time which requires a conceptual apparatus apt for the purpose. Thus, Sinclair operates with the terms 'the idiom principle' and 'the open-choice principle'.

The distinction between them is determined by the nature of the choice at any moment in text: whether the choice of the lexical item to follow is conditioned (by previous exposure) and therefore predictable, or is made independently. These are two modes of text interpretation/production which switch into each other. “Trusting” the text and looking at its real-time moment by moment progression, Sinclair arrives at a model which accommodates a great deal of variability and at the same time acknowledges conditioning of choices at very abstract levels. This line of thinking is further developed in Linear Unit Grammar (Sinclair and Mauranen 2006). The notion of a “prefabricated” item is much less flexible. In fact, it seems closer to the 'stock of phrases' interpretation mentioned in Chapter1.

This difference in conceptualisation is reflected in the methods Wray and Sinclair preferentially use in identifying formulaic sequences and units of meaning respectively. For Wray intuitive judgments of native speakers are insightful (Wray 2002, 2008, 2009), and she works at developing “diagnostic criteria for assessing intuitive judgments about formulaicity” (Wray 2008: 116; first proposed in Wray and Namba 2003). This view rests on the assumption that native speakers are able to identify a formulaic sequence when they see it. In particular, she states that “[w]e should anticipate that formulaic sequences shared across a speech community can be reliably identified by most native speakers, provided they know what they are being asked to look for” (Wray 2008: 107). In contrast, in the framework of the idiom principle, the conditioning of the choice may not be available to intuition until recurrence of the pattern is revealed through a corpus search and vertical or paradigmatic arrangement of usage instances in concordance lines (see Sections 2.1, 2.2 and 2.6.4). So not only a native speaker judge but even the hearer or the speaker herself may not be able to say which stretches of language are produced on the idiom principle. For this reason, while frequency of co-occurrence has its problems, which are quite legitimately pointed out in Wray 2002 (28-31) and Wray 2009 (36-37), it can sometimes be the only indication of operation on the idiom principle and underlying syntagmatic associations.

In fact it is possible that for Wray formulaic language constitutes part of declarative knowledge. This can be seen in the assumption that a language user is able to choose between formulaicity and novelty, that is, to make a conscious decision whether to rely on a formulaic sequence or to “abandon formulaicity in order to release new meaning” (Wray 2008: 48). Further, Wray argues that formulaic language is in many ways advantageous to novel constructions because “[t]he more novel our output is for the hearer, the more likely it is to be misunderstood” (2002: 94). To illustrate the idea, Wray gives an example of army commands, which have low potential for misunderstanding or slow reactions. Another example is using a formulaic expression *Excuse me* when you want to apologise for having to leave without attracting too much attention instead of a more novel construction "*It's time for me to live now*" (Wray 2002: 95). In other words, what these examples show is that if we want to achieve quick and reliable interpretation of our utterance on the part of the hearer, we had better use formulaic sequences.

In contrast, the idiom principle is connected with the idea of syntagmatic prospection. The hearer half-expects what he is going to hear and the expectation is then confirmed or overthrown, the latter may of course result in confusion (or an ironic effect):

> [Structure] allows the reader or listener to *prospect* ahead and make informed guesses about what is likely to come. So instead of the listener hearing a burst of sound and then trying to work out what it means, the listener will be half-expecting one of a few options, and will only need to confirm which one it is. [...] The frequency of phraseological patterns is another origin of expectations, because prospection arises from experience, and so common patterns will be more securely expected than rare ones. (Sinclair and Mauranen 2006: 135-137)

In other words, while Wray suggests that formulaic language is easier to process because the form-meaning mapping of a formulaic sequence is more reliable and its decoding is more likely to be communicatively successful, Sinclair and Mauranen talk about the interdependence between the frequency of a pattern, or the likelihood of it being familiar to the hearer, and its prospection facilitating effect. The prospection effect takes place when the pattern used by the speaker is familiar to the hearer, i.e. it does not have to be established as formulaic in the language as a whole.

2.11. Psycholinguistic reality of a unit of meaning: a summary

By now the phrase 'psycholinguistic reality' has already been used several times in relation to the concept of a unit of meaning. Yet, what does psycholinguistic reality mean and do all patterns identified in language use have to be psycholinguistically real in order to be valid descriptions of language?

As explained in Chapter 1, the focus of this study is on L2 users' phraseological competence. It is a widely shared view that an ability to acquire and use multi-word units is determined by propensity for holistic processing, which has also been hypothesised to be lacking in L2s – an assumption which this study seeks to examine closely. Given this focus, it was important to select a psycholinguistically real model of a multi-word unit, i.e. the one which takes cognitive constraints into account and is able to reflect how a multi-word unit is represented in the mind. In other words since this study is about acquisition, processing and use of multi-word units rather than for example social construction of meaning influencing language patterning or cultural differences in language structures expressing politeness, it attempts to identify those language patterns which are determined by the properties of cognitive processing.

The model of a unit of meaning is postulated to be psycholinguistically real for several reasons. First of all, the idiom principle is hypothesised to be a psycholinguistic mechanism of language production by co-selection. It is the idiom principle which launches meaning-shift and by virtue of which a combination of words becomes a unit of meaning (or a meaning-shift unit). In other words, it is the fact that they are co-selected, produced holistically, which makes the words lose their 'core' meanings or delexicalise and acquire a larger shared meaning. Second, lexical items are hypothesised to be represented in the mind according to their meanings rather than forms (see Section 2.6.4), therefore, the completeness and independence of meaning in a unit of meaning should be a critical feature for psycholinguistic representation. Third, the optional components of a unit of meaning are defined as syntagmatic associations, verbatim in the case of a colocation, and abstracted, in the case of a semantic preference and colligation. Further on, in Sections 3.5 and 3.6, it will be suggested that cognitively semantic preference and colligation come about as approximated collocations due to frequency effects and the superiority of memory for meaning over memory for form. Taking all this into account, a unit of meaning as a pattern is a product of psycholinguistic processing and can be explained by domain-general cognitive processes[27] such as sequential processing, which lies behind the emergence of syntagmatic association between items used together, categorisation which is presumably responsible for the emergence of the categories of semantic preference and colligation, supremacy of meaning for the memory system and sensitivity to statistical information about language regularities - all these realities of psycholinguistic processing conspire in such a way that leads to the emergence of a unit of meaning in the form we are able to observe using corpus linguistic data.

At the same time, if we distinguish between a communal and a cognitive level in language (see e.g. Beckner et al. 2009 or Ellis 2011), or a cognitive and a macrosocial perspective on language (Mauranen 2012), it can be said that a unit of meaning exists or has a

[27] See e.g. Bybee 2010 on the role of domain-general cognitive processes in the emergence of language structure.

reality at both levels. So for example, a general corpus such as the BNC enables observations of language use at the communal level, a corpus of a language user's written production, such as the ones used in this study, is a window on the cognitive level, with the reservation that it needs to be complemented with other kinds of data to make psycholinguistic observations possible. There seems to be a complex interaction between these two levels. For example I might like saying "*X drives me crazy"*, but other language users may have other preferences: *X drives me crazy / mad / wild / up the wall / nuts*.[28] Each of them might have a favourite collocation, cognitively, a verbatim association, while at the communal level such 'cognitive collocations' cumulatively form a fuzzier category of semantic preference.

It is also clear that a unit of meaning is not the only unit which exists in language. Different kinds of units can be identified from a variety of different perspectives and using different techniques. What we have at our disposal as linguists, just as ordinary language users do, is a stretch of linear language text. This stretch of text can be divided into different kinds of units or patterns. The defining feature of a language pattern is that, on the one hand, it permits a certain degree of predictability or prospection and, on the other, contains certain dependent choices. The choice(s) within the pattern can be conditioned by different factors, cognitive as well as social (Section 2.8.2 calls such conditioning effect "dependent choices at different levels"). For example units of meaning and Linear Unit Grammar chunks (Sinclair and Mauranen 2006) are mostly cognitively determined, semantic sequences and e.g. Create-a-Research-Space (CARS) model in research articles (Swales 1990) are mostly socially determined patterns. At the same time there seems to be a complex interaction of different factors in different patterns, for example the socially determined principle of spoken interaction such as "keep talking" (Biber et al. 1999: 1067) adds to the pressures of online processing affecting the length and structure of the units produced. There are other factors as well, for example the fact that language in use is linear, which might be called a 'linguistic factor', definitely has a serious impact on how speech or written text is structured, and as a result on the emerging patterns determined largely by this factor. Different kinds of units can be multiple-embedded and overlapping which creates a complex interaction between them but at the same time presumably increases the possibilities of prospection for the reader/hearer ensuring comprehension.

[28] Cf. Bybee's (2010) analysis of this phrase.

2.12. Conclusion

In this chapter I have argued that Sinclair's unit of meaning, an entirely corpus-driven concept, is distinct from other notions of a multi-word unit in that it is conceptualised as a product of the idiom principle which seems to amount to a psycholinguistic mechanism of language production based on making use of implicit memory and syntagmatic association. A sequence of linguistic elements becomes a unit of meaning – a unit which serves to communicate just one holistic meaning – by virtue of being produced on the idiom principle because where there is one choice, there is one meaning. Any sequence which is produced on the idiom principle undergoes a meaning-shift giving way to a larger holistic meaning since words which are co-selected are always delexicalised even if to a smallest degree. Therefore, a co-selected sequence is a new form-meaning pairing which has nothing to do with the form-meaning pairings of its constituent elements, single words. The model of a unit of meaning has the advantages of integrating syntagmatic and paradigmatic axes of patterning and allowing for formal variation inside the limits of one invariable meaning, its semantic prosody.

One of the major arguments of this chapter is that it is useful to see Sinclair's conceptualisation of lexis and meaning as a system where all the elements – collocation, unit of meaning, semantic prosody, idiom principle – are inextricably intertwined. Semantic prosody is therefore seen as a component of a unit of meaning. It seems that the quest for semantic prosody and trying to get to the bottom of it has led to the discovery of a whole collection of new concepts and terms, such as Louw's 'logical' variant of semantic prosody and attitudinal preference suggested as a term (Hunston 2007) to describe often detected negative and positive associations of a lexical item. All of these terms deserve their own place in linguistic theory. But it is contended here that the concept of semantic prosody is best restricted to the unit of meaning. It was also argued that as a component of a unit of meaning, semantic prosody is not inherently evaluative, evaluation being just one functional type of possible semantic prosodies. As such it should be clearly separated from neighbouring concepts like connotation, semantic preference and attitudinal preference. Diachronic perspective is not relevant when we talk about semantic prosody as a communicative purpose of a unit of meaning since idiom principle is a synchronic mechanism. It is also suggested that semantic prosody is not hidden. What can be hidden from our intuition instead is the pattern, or, to be more specific, the recognition of a larger unit by its individual parts especially if they are substantially delexicalised.

Further, it has been argued that claims about lexical patterning must take meaning into account in order to be psycholinguistically plausible since it is reasonable to assume that representations in the mind are organised according to meaning rather than according to form. Sinclair's unit of meaning, which is distinct from Firth's collocation seen as a way of interpreting the meaning of a word, is a good candidate for psycholinguistic reality since it is characterised by an independent meaning of its own. In other words, it is not only frequency of co-occurrence which affects the representation in the mind, but also the wholeness of the meaning the co-occurring sequence communicates. The notion of co-occurring meanings or co-occurrence at any other higher levels of abstraction might be beyond psycholinguistic consideration.

Huge variability in language patterning may convince us that the lexicon is actually "empty" (Sinclair 1996b), in other words it may be more fruitful to concentrate on the principles of this patterning and the mechanisms which underlie them rather than fixed phrases per se. Phraseology is at the heart of language, but at the heart of phraseology is variability. The terms 'phraseological tendency of language' and 'syntagmatic prospection' emphasise this emergent property of lexical patterning and foreground the process rather than the product, the prefabricated store of phrases.

3. Second language acquisition and use of multi-word units

Multi-word units in second language acquisition and use is a topic which attracts attention of many scholars for different reasons - pedagogical, lexicographical, theoretical - and the number of studies devoted to it is truly large. The aim of this chapter is therefore not to give a comprehensive overview of all the studies carried out on the topic but to summarise what seem to be the mainstream views on the ability of second language users to operate with multi-word units, that is, as many see it, on the idiom principle.

The common conclusion most researchers come to is that phraseological competence constitutes a major problem for second language learners even at advanced level of proficiency. This view will be explained in Section 3.1. Section 3.2 will present some examples of studies which share this view. A natural question arising from the findings of these studies is why handling multi-word units is so difficult for L2 learners. Alison Wray offers a psycholinguistic explanation of this phenomenon, which will be presented and discussed in Section 3.3. Since this explanation is not intuitively entirely satisfying, in Section 3.4, I will go back to the studies described in Section 3.2 and revisit their methodological and conceptual assumptions which could have influenced the conclusions about the phraseological problem they inevitably come to. In Section 3.5, an alternative explanation of the commonly observed shortcomings of L2 phraseological competence will be provided. It will be argued that certain changes L2 learners introduce into standard forms of multi-word units can be explained through a process of approximation. It will also be proposed that approximation is a natural process which works inside a unit of meaning. Section 3.6 will suggest a cognitive basis for this explanation.

But before I start, it is important to make a note on the use of terminology in this chapter. As it was explained in Chapter 2, this study is based on Sinclair's theoretical framework in which multi-word unit patterning a language exhibits is seen as a phraseological tendency or phenomenon. This makes it more appropriate to talk about syntagmatic prospection and syntagmatic organisation of language rather than a stock of prefabricated expressions which are different from the rest of language. Yet, this is not the mainstream view; therefore, I cannot discuss other studies on the phraseology of L2 learners in these terms. As pointed out in Section 2.1, there is an impressive number of terms which are used to describe the phraseological phenomenon, and learner language studies are no exception to this. It will be argued in Section 3.4 that lack of terminological agreement is unfortunate, but "disentangling the phraseological web", as Granger and Paquot (2008) put it, is not the purpose of this study. The assumption is made that patterns described by different authors are anyway somehow related to the phraseological tendency of language, even if partly. Therefore, when describing other studies, I use the preferred term of the author(s). When I need a generic term to refer to all kinds of phraseological patterning, I use the terms multi-word units or phraseological units. In case of fuzzier patterns, when the specific “units” are not necessarily identifiable, I use the term patterning. The term lexical item implies that it can be applied to both a single word and a multi-word unit (see Section 2.3). I only use Sinclair’s terms - a unit of meaning, an extended unit of meaning or a meaning-shift unit - when the patterning in question was analysed in terms of a unit of meaning and a unit of meaning was established. While I use collocation in a strict sense, either as a component of a unit of meaning which is characterised by verbatim association or a unit of meaning which has the structure of a collocation; in research literature it is usually defined as a statistical or a psychological association between two words.

The term second language (L2) users is superordinate in relation to L2 learners, that is, although L2 users can be represented by L2 leaners, they can also be English as a lingua franca (ELF) speakers. The term L2 learners is used when it is clear that it is this group of L2 users which is referred to, in the work cited, for example. I use the term L2 users for the most part because I am interested in L2 processing which unites ELF users and L2 learners (see Mauranen 2011).

3.1. Phraseology seen as a major problem for language learners

As discussed in Chapter 2, the idiom principle seems to be one of the major principles of language production. For example, it can explain more than half of running text in English, whether written or spoken (Erman and Warren 2000). Yet, the mainstream research in Second Language Acquisition (SLA) seeks to emphasize that the idiom principle facilitates language production and comprehension only for those language users who acquired the language in question as their mother tongue. Collocations and multi-word units (MWUs) at large are usually reported to be the major stumbling block for second language learners. Moreover, it is very often claimed that idiomaticity is one of the features that distinguishes native speaker (NS) and non-native speaker (NNS) language use.

Pawley and Syder (1983) posed nativelike selection and nativelike fluency as two puzzles for linguistic theory more than thirty years ago. They suggested that “lexicalized sentence stems” which exist in addition to productive rules of grammar and enjoy the special status of a lexical item being retrieved holistically from memory are able to explain both of the puzzles. With reference to second language acquisition, they point out that:

> It is a characteristic error of the language learner to assume that an element in the expression may be varied according to a phrase structure or transformational rule of some generality, when in fact the variation (if any) allowed in nativelike usage is much more restricted. The result, very often, is an utterance that is grammatical but unidiomatic e.g. *You are pulling my legs* (in the sense of deceiving me). *John has a thigh-ache*, and *I intend to teach that rascal some good lessons he will never forget!* (Pawley and Syder 1983: 215)

Their observation has found support in subsequent studies investigating English as a Foreign Language (EFL) or English as a Second Language (ESL) learners' mastery of multi-word sequences. The common conclusion researchers come to is that language learners do not make full use of phraseological expressions in spite of the benefits of sounding nativelike and gaining fluency they provide.

3.2. Learner language research: NS vs. NNS

An overwhelming proportion of the studies on learner use of MWUs are corpus-based. We will look at some of the most prominent of them to see how they arrive at the conclusion that MWUs present the biggest problem for second language acquisition and use.

Howarth (1998) looked at lexical collocations (verb + noun complement) in NS corpora (social-science texts from LOB corpus[29] and some additional NS expert academic texts) and learner corpora compiled for the purpose and comprised of essays written by NNS students attending a 1-year Master's course. In line with the general trend, he found that "native speakers employ about 50 percent more restricted collocations and idioms (of a particular structural pattern) than learners do, on average" (177). In addition to this main finding, I would like to emphasise two points in Howarth's research. First, it is interesting that in his study NSs used a number, although a small one, of forms that were deviant from standard collocational forms. What is more important is that the deviations could be divided into two main types: grammatical modification and lexical substitution (Howarth 1998: 171). Howarth found the two categories useful in analysing learner data as well. Second, he laid particular stress on the fact that in his analysis of non-native speaker collocations there was no predetermined set of collocations. Neither did the collocations identified have to match the

[29] The Lancaster-Oslo/Bergen Corpus.

collocations from the respective NS corpus since he analysed all non-native speaker texts manually and extracted all the collocations that matched the criteria he adopted for identifying lexical collocations. However, as it seems, it escaped from his view that exactly because he analysed the writing samples manually, he extracted a predetermined set of collocations since they had to conform to his idea of nativelikeness. In other words, all the "non-nativelike" but recurrent combinations in learner writing were not retrieved.

Granger (1998) worked with the French subcorpus of ICLE[30] consisting of argumentative essays and a comparable NS corpus. She hypothesized that "learners would make much greater use of what Sinclair (1987: 319) calls the 'open-choice' principle than native speakers, who have been found to operate primarily according to the 'idiom principle" (Granger 1998: 146). To test the hypothesis she compared the use of collocations and formulae in the two corpora. As a result she found that in relation to native speaker usage NNS generally underused amplifiers and especially the amplifier *highly*, but at the same time overused what seemed to be their favourite amplifiers *completely* and *totally* and "the all-round intensifier *very"* (Granger 2009: 22), which was taken to suggest that "learners seem to use amplifiers more as building bricks than as part of prefabricated sections" (Granger 1998: 151). What concerns "sentence-builders", learners "massively overused the active structure" (Granger 1998: 155) and seemed to "cling on' to certain phrases and expressions which they feel confident in using" such as the phrases *I think that, I would say that* (Granger 1998:156). Granger concluded that "learners' phraseological skills are severely limited: they use too few native-like prefabs and too many foreign-sounding ones" (Granger 1998: 158).

Paquot (2008) compared MWUs used for the purpose of exemplification in learner writing and native speaker writing using NNS argumentative essays of ICLE and American students' essays of the same kind collected in LOCNESS. The essays in both corpora are written on the topics such as "death penalty", "euthanasia", "crime does not pay", "money is the root of evil" (Paquot 2008: 104). The items under analysis are "word-like" units, as Paquot calls them, *for example* and *for instance*, and "collocations and frames with the noun *example* and the verbs *illustrate* and *exemplify*" (Paquot 2008: 107). "The striking differences" (Paquot 2008: 108) between NS and NNS usage Paquot reveals in her study include learners' overuse of *for example* and *for instance* and underuse of the above mentioned frames, except for the active structures with the verb *illustrate* and the phrase *To*

[30] ICLE stands for the International Corpus of Learner English, which is comprised of argumentative essays collected from high intermediate to advanced learners of English coming from different L1 backgrounds. See http://www.uclouvain.be/en-cecl-icle.html

illustrate this..., which they overuse. Paquot suggests that learners tend to cling on to the fixed word-like units for two main reasons: probable transfer effect, since the learners' first languages have direct equivalents of *for example* and *for instance,* and their overemphasis in teaching.

Paquot also points out that "these word-like units are repeatedly used when they are unnecessary, redundant or even when other rhetorical functions should be made explicit" and gives the following example from learner writing: "*The mob **for instance** is a very good **example"*** (2008: 110, original emphasis). Yet a quick look at, say, BNC yields a very similar case from its Academic prose subsection: *...materials produced and developed at Dudly Teachers' centre, for instance, is an excellent example.* It is possible that Paquot's conclusion is influenced by an overly critical and unforgiving attitude to language learners who by definition are supposed to have worse language skills than native speakers, an attitude which is widespread in the field.

With reference to transfer-related effects, Paquot further draws her attention to the French learner population's "massive overuse" of the phrase *let us/let's take the example of,* which, as she insists, is most probably caused by the existence of an equivalent expression in French. In her view the usage is inappropriate since "[a]cademic writing is a genre characterized by high degrees of formality and detachment". However, if we remember the topics of the essays which comprise the corpus, it becomes questionable whether the fact that the authors of the essays are university students is sufficient to regard their writing as academic. It seems that essentially academic texts are usually written for a different audience and with a different purpose. In contrast, an essay is very often used as a practical exercise in learning to write in one's mother tongue or a foreign language. For example, Mauranen (2011) suggests that 'composition' is one of the "pedagogical genres that are specific to educational settings only" (Mauranen 2011: 158). Therefore, it is arguably hard to talk about "lack of register awareness" (Gilquin et al. 2007: 319) in EAP learner writing on the basis of ICLE corpus (see Ädel 2006 for some other critical comments about ICLE).

Granger and Paquot (2009) explore lexical verbs in academic discourse. They compare ICLE argumentative essays with NS expert academic writing and find that learners tend to underuse the lemmas from Coxhead's Academic Word List (AWL, Coxhead 2000) or, with even a higher proportion, from Paquot's Academic Keyword List (AKL, Paquot

2007, 2010), but overuse lemmas from General Service List (GSL, West 1953).[31] Yet, they also point out that looking at lemmas instead of word forms may distort the picture in certain important ways, in contrast, word form patterns can provide more detailed descriptions of the "behaviour" of several verbs in learner and expert writing. For example, learners do not seem to have difficulties with the verb CONCLUDE if we only look at the lemma frequencies. However, it turns out that the infinitive form is overused while the forms *concludes* and *concluded* are underused. In practice it means that while learners prefer to finish their essays with *To conclude...,* expert writers come up with a range of ways they present their concluding remarks: *Finally, the chapter concludes by..., He concludes that the effectiveness..., It is reasonable to conclude from this that..., We may conclude that,* ...(Granger and Paquot 2009: 209). A counter-argument could be that essays are not written in chapters or by multiple authors and it is not typical to refer to external literature sources in essays either, especially if the essay is written in the format of a timed exam which is the case for the majority of learner essays in the ICLE corpus. The authors admit the comparability problem (see Granger and Paquot 2009: 197, 208), but arguably dismiss it too easily, stating their findings as follows:

> The first [finding] is that EFL learners significantly underuse the majority of 'academic verbs', that is, verbs like *include, report* or *relate,* that express rhetorical functions at the heart of academic writing, and instead tend to resort to 'conversational verbs', that is, verbs like *think* or *like,* that are characteristic of informal speech. The second is that when learners use academic verbs, they tend to restrict themselves to a very limited range of patterns, which contrasts sharply with the rich patterning that characterizes expert writing. (Granger and Paquot 2009: 210)

The two authoritative volumes on phraseology or formulaic language, one edited by Norbert Schmitt (2004) and the other by Fanny Meunier and Sylviane Granger (2008), take the problematic nature of collocation for language learners as one of their fundamental initial assumptions. In the introduction to the first volume Norbert Schmitt and Ronald Carter write:

> As learners' proficiency improves, there is the reasonable expectation of language which is more accurate and appropriate. In natives, this is achieved to a large extent through the use of formulaic sequences. Unfortunately, the formulaic language of L2 learners tends to lag behind other linguistic aspects (Irujo, 1993). This may be partly due to a lack of rich input: Irujo (l986) suggests that idioms are often left out of

[31] They found that 23/50 most underused verb lemmas belong to AWL (Coxhead 2000), (or 44/50 to AKL (Paquot 2007), 45/50 most overused come from GSL.

> speech addressed to L2 learners. Learners also seem to avoid the use of idiomatic language (Kellerman, 1978), although this may have more to do with the degree of Ll-L2 similarity than any intrinsic difficulty (Laufer and Eliasson, 1993; Laufer, 2000; Vihman, 1982:272). There is also the tendency to stick with familiar and 'safe' sequences which the learners feel confident in using (Granger, 1998), although De Cock (2000) found that some formulaic sequences were overused, some underused, and others simply misused by nonnatives when compared to native norms. (Schmitt and Carter 2004: 13)

Completely in line with these observations, Granger and Meunier state in their concluding remarks that:

> Linguistic analysis has amply demonstrated the patterned nature of language, both lexically and grammatically, stressed the pervasiveness of phraseology in oral and written communication, and the difficulties the learners have in mastering native-like phraseology. (Granger and Meunier 2008: 247)

In other words, most of the corpus-based studies on learner use of MWUs seek to show that, to use Granger's description, "learners' phraseological skills are severely limited" and in need of "remedial" pedagogy (Granger 2009: 22). It is also felt that the most probable out of the potential reasons for the difficulties found is that, as Kjellmer put it, learners' "building material is individual bricks rather than prefabricated sections" (Kjellmer 1991: 124) or, as Granger hypothesized, learners in contrast to native speakers operate on the open choice principle rather than on the idiom principle. Therefore, learners' errors are due to the necessity to construct a phrase each time from separate words.

3.3. Wray's psycholinguistic explanation of the problem

Alison Wray has proposed a psycholinguistic explanation as to why learners cannot take advantage of the idiom principle to the extent natural to native speakers. In the introduction to her 2002 book, Wray writes that she has always been fascinated by the disproportion between the ease with which native speaker children acquire formulaic sequences by picking them up from adults' talk and the problems language learners constantly experience with the same sequences: they just won't get them right. She suggests that the theoretical solution to this puzzle resides in the storage mechanisms of the mental lexicon which are different for first and second language acquisition processes. Wray argues that second language learners are not sensitive to formulaic sequences but instead tend to focus on individual words, acquiring and storing separate words rather than holistic phrases as a result (Wray 2002), which is in line with the conclusions the researchers in SLA and learner corpus studies discussed above come to.

Wray starts her discussion of the differences that have to be introduced into the first language model of formulaicity when we are dealing with second language learners by referring to Yorio's (1989) study of idiomaticity and second language proficiency. Yorio observes that advanced ESL students attempt to use formulaic sequences in their writing but end up with a considerable amount of errors, for example: *take advantages of; are to blamed for; those mention above; being taking care of; a friend of her; make a great job; on the meantime; with my own experience; put more attention to* (Yorio 1989: 62-63). As a reason for the occurrence of such errors, he suggests that "these expressions are not simply memorized or taken in as wholes, but ... are subject to whatever interlanguage rules the learner is operating under" (Yorio 1989: 62). Wray speculates that since the attempted sequences are close to those used by native speakers, it is plausible to assume that the students have been exposed to the "correct versions", yet they still got them wrong, which is puzzling (Wray 2002: 199). Wray analyses the case further and offers two alternative interpretations: "The strings could have been memorized, but incorrectly, because the interlanguage grammar edited the forms to something consistent with its expectations. Alternatively, they could have been correctly memorized but edited as part of the production process" (Wray 2002: 200). However, she goes on, in both cases we should be able to see how the learner's interlanguage grammar worked either at the time of memorising the sequence or at the time of its production: "in the former case, the errors would be consistent with the grammar at the time of learning and, in the latter, with the grammar at the time of use" (Wray 2002: 200). To solve the problem, Wray refers to Granger (1998) who argues that language learners are unlikely to be able to analyse a formulaic sequence grammatically:

> ...there does not seem to be a direct line from prefabs to creative language, or to use Sinclair's (1987) terms, from the idiom principle to the open choice principle. It would thus be a foolhardy gamble to believe that it is enough to expose L2 learners to prefabs and the grammar will take care of itself. (Granger 1998: 158)

Taking this important claim into account (which is at any rate somehow contradictory to the idea that language learners are not able to memorise formulaic sequences holistically), Wray suggests that possibly learners do engage in an analysis of what appears to be a holistic string but in a different way: they break it down into words without paying any attention to the grammatical information it contained:

> The result would be that the learner had a store of the *words* which had occurred in the formulaic sequence, but none of the detailed grammatical (particularly morphological) information about how they combined. In reconstructing them, the correct words would, therefore, be conjoined according to the current interlanguage rules. (Wray 2002: 200)

This logical conclusion is going to be the central tenet of the model of the second language mental lexicon Wray develops later in the book. As mentioned earlier, the model is based on the differences between first and second language acquisition of formulaic sequences:

> Where the first language learner starts with large and complex strings, and never breaks them down any more than necessary, the post-childhood second language learner is starting with small units and trying to build them up. Phrases and clauses may be what learners encounter in their input material, but what they notice and deal with are words and how they can be glued together. The result is that the classroom learner homes in on the individual words, and throws away all the really important information, namely, what they occurred with. (Wray 2002: 206)

So, for example a native speaker would naturally treat a collocation like *major catastrophe* as a formulaic sequence and would leave it unanalysed by default in accordance with the needs-only analysis principle (NOA, Wray 2002: 130). In such a way the sequence would be memorised and stored holistically contributing to the puzzle of nativelike selection. Second language learner's processing is predicted to be different:

> In contrast [to NS], the adult language learner, on encountering *major catastrophe* would break it down into a word meaning 'big' and a word meaning 'disaster' and store the words separately, without any information about the fact that they went together. When the need arose in the future to express the idea again, they would have no memory of *major catastrophe* as the pairing originally encountered, and any pairing of words with the right meaning would seem equally possible: *major, big, large, important, considerable*, and so on, with *catastrophe, disaster, calamity, mishap, tragedy*, and the like. Some of these sound nativelike others do not, and the learner would have no way of knowing which were which. (Wray 2002: 209)

In the end we have a quite gloomy picture of the fate of a language learner:

> ...the non-native speaker, however accurate in grammar and knowledgeable at the level of words, would always be a potential victim of that lesser store of formulaic sequences. There would be situations in which a native speaker would call on an idiomatic prefabricated expression, while the non-native had to create one, an activity

> both more labour intensive and more risky. As a result, language production and comprehension might always feel more effortful than in the native language. (Wray 2002: 210)

3.4. Revisiting the approach of learner language research

Wray's explanation of the SLA puzzle may seem to be plausible. However, if we return to the very beginning of the discussion, we might notice several problems which could have affected this way of thinking. First, we do not yet seem to have reached an agreement on what phraseology is and what counts as a MWU. The second is basically a methodological problem: the kind of evidence which is used to argue that NNSs do not take advantage of the idiom principle may not be valid for making such claims. The fact that NNS production does not contain enough examples of phraseological units identified in NS production does not yet automatically mean that it does not contain phraseological units at all. Part of this problem of course stems from the previous one: when phraseology is conceptualised as a stock of phrases which consist of several words rather than just one, it is tempting to assume that to operate on the idiom principle means to have this stock available and that therefore it should be possible to determine whether a language user is operating on the idiom principle, by counting how many phrases s/he is using out of this stock. This approach might be misleading. I will now discuss these two problems in more detail.

The different approaches to phraseology were discussed in Chapter 2. It has often been stressed in the literature that there is no single uniform definition of phraseology, i.e. different authors may count different stretches of language as MWUs. This may be seen as an acute problem since lack of agreement precludes any direct comparison between studies. At the same time, it is natural that the definition might differ depending on the approach we take, whether we operationalise a MWU through, for example, corpus linguistic or psycholinguistic methodology. This can be beneficial for linguistic theory, as different approaches may inform each other, open new facets in the object of research and reveal weak points in the accepted "truths".

Erman (2009), for one, stresses the importance of working out "an approach to collocations that would benefit learners" (344). She argues that counting only the "'restricted' or problematic" (Erman 2009: 330) collocations is not useful from the learner perspective and that "the focus on phraseological units in the literature has over-shadowed the fact that there is an abundance of multiword expressions that, although they will not pass the phraseological grid, nevertheless have specific, unitary meanings, connected to specific cultural frames, which have to be learnt" (Erman 2009: 344). Erman herself defines collocations as "composite structures with unitary meanings" (Erman 2009: 331). For her analysis, she selects two categories of collocations - Verb+Noun and Adj+Noun - and divides them into three groups according to their meanings: "Lexical Functions", "socio-culturally motivated" collocations and topic-induced frames. While the first two groups can be convincingly claimed to "have lexical status, i.e., have specific unitary meanings just like single words and are presumably stored holistically, or at least easily retrieved, one member calling up the next through associative networks" (Erman 2009: 331-332), the last group of collocations "have not (yet) reached" this status because they evolve as key to the texts under examination. Since her definitions are based on the functions which collocations play in language, that is, on their meanings, she does not need frequency information to identify them. So she extracts all Verb+Noun and Adj+Noun combinations from her data of non-native writing samples, identifies collocations according to her meaning-based criteria and measures the proportion of these collocations out of the total number of Verb+Noun and Adj+Noun combinations occurring in the data instead of directly comparing MWUs found in NS writing to MWUs in NNS writing. According to her estimates 39.8% of all word combinations in her learner data are collocations suggesting that "[t]he idiom principle is the default principle in all language production for learners and native speakers alike (Erman 2009: 341) even though the respective figure for NS data is 60.2% and the collocations learners used did not necessarily match NS preferences.

Nesselhauf (2005), who manually extracted all verb-noun collocations from the German subcorpus of ICLE (i.e. she did not have a predefined list of MWUs either), also rejects the idea that idiom principle may not be available to non-native speakers, instead she prefers to talk about "the degree to which second language learners rely on chunks and to which they creatively combine individual words in language production" (247):

> While it has been shown that learners use an overall smaller number of prefabricated units than native speakers (e.g. Granger 1998; Kaszubski 2000), the claim that they only use very few prefabricated units can be refuted on the basis of the present analysis. Learners did use a large number of native-like collocations (cf. Section 3.1), and although it cannot be assumed that all of them were stored and produced as chunks, it is improbable that the majority of them was creatively combined by the learner in a way that coincided with native speaker collocations. (Nesselhauf 2005: 247).

Nesselhauf also points out that even the cases of inappropriate usage of existing collocations and usage of non-nativelike word combinations cannot be taken as indisputable evidence of the learners' preference for the open-choice principle (see Nesselhauf 2005: 247-248). For example, a collocation in learner's repertoire may be affected by L1 transfer, yet it may very well be stored and retrieved as whole. Moreover, such slightly incorrect collocations can also be picked up from other learners in a nativelike formulaic fashion.

The two studies by Nesselhauf (2005) and Erman (2009) are good examples of how it is possible to arrive at almost opposite to the mainstream conclusions about the availability of the idiom principle to second language learners by taking a little bit different perspective on phraseology and changing the definition of a MWU together with its operationalisation.

The second problem of studies on second language use is that, as we have seen from the examples of studies done in the field, the conclusion about the lack of idiomaticity in second language use has been made almost solely on the explicit or implicit comparison of NS vs. NNS production where the idiomaticity of the former is assumed by default. If we look at the studies on second language use of MWUs, especially corpus-based, more closely, it will be evident that they tend to fall into a certain pattern. To some extent this is natural since the pattern suggests itself as an application of the methodology of Contrastive Interlanguage Analysis (CIA) proposed by Granger (1996). The CIA method includes two types of comparison: first NNS data is compared to NS data in order to "highlight a range of features of non-nativeness in learner writing and speech, i.e. not only errors but also instances of under- and overrepresentation of phrases and structures" (Granger 2002: 12). Then NNSs from different mother tongue backgrounds are compared between each other, and this second comparison makes the analysis particularly suited for the investigation of transfer. For example, if as a result of the first type of comparison we find that Finnish learners of English use a non-nativelike phrase DISCUSS *about,* and its frequency is significant enough to claim that it is not an idiosyncrasy but is spread among all Finnish speakers of English, we cannot yet claim that it appeared in the interlanguage of Finns as a result of transfer of a Finnish structure, although this structure indeed seems to be capable of provoking such a mistake. In order to test the hypothesis of a possible transfer effect we need to find out whether speakers of other mother tongue backgrounds use the phrase or whether it is restricted only to Finnish speakers of English. The logic of this argument is based on Jarvis's criteria for identification of L1's cross-linguistic influence in an interlanguage (Jarvis 2000). It is easy to see that the ICLE corpus serves the purpose extraordinarily well, perhaps too well, since it makes learner corpus research studies predictable. The design of ICLE presupposes studies of a 'compare and contrast' type. We hypothesize that a certain feature would be represented differently in NS and NNS data and then we look whether it is represented differently in learner subcorpora as well. One of the potential problems is that the differences 'uncovered' in this way may in reality stem from unrelated factors, let alone that if we are specifically looking for differences, it is quite rare that we do not find them, especially if we are dealing with overuse and underuse counts.[32]

One of these factors is the question of comparability. Among the studies reported earlier there are studies which compare ICLE with LOCNESS (or part of it) consisting of American students' as well as British students' and pupils' essays.[33] This comparison raises a question whether an essay written by an American or a British student should be considered a model for an EFL learner since there may be important cultural differences in what is considered a good essay that may have to be taken into account (see e.g. Leech 1998: xix). The comparison between ICLE and a corpus of expert academic writing, which is plausible as a model writing at least for those who are aspiring to become academics, in its turn is questionable because of the differences in text-type. It is generally accepted that language varies with genre, register and format, which precludes, as one might feel, any comparison between an essay and an article or a chapter in an academic volume. Constructions employed differ considerably even within one genre of academic prose, the research article, depending on the discipline (Hiltunen 2010). It should be noted that the authors of the study themselves admit the comparability problems:[34]

> An important caveat, however, is that the two corpora are not fully comparable. Expert texts are expository in nature, that is, they are topic-oriented (cf. Britton, 1994) and rely on the comprehension of general concepts (cf. Werlich, 1976), while argumentative essays start 'from the assumption that the receiver's belief must be changed' (Gramley & Pätzold, 1992: 193). In addition, expert texts are discipline-specific while learners' essays discuss a range of general topics such as feminism, the impact of television, drugs, etc. (Granger and Paquot 2009: 197).

[32] For some reason, if we conclude that learners underuse a certain feature, we do not make a logical connection that perhaps native speakers overuse it, but when learners overuse a certain feature, it is a "lexical teddy bear" or an "island of reliability". Learners seem to be caught in a no-win situation.

[33] For more information see http://www.uclouvain.be/en-cecl-locness.html

[34] See Ädel 2006: 201-203 for a thorough discussion of other comparability problems at play within the design of the ICLE corpus and Ädel 2006: 205-208 for a review of different viewpoints on the question of why a corpus of American students' writing or a corpus of expert academic writing may not be particularly suitable as control corpora for a corpus of learner essays.

Another question which is perhaps less evident is whether we can talk about *an interlanguage* of a learner population especially when grouped on the basis of the mother tongue background. By dividing the learners of English into groups in this way, we a priori assert L1 influence as the major factor in second language acquisition and use and are blinding ourselves to any other factors that may also play an important role. While there is ample evidence that effects of transfer take place, by assuming their superiority from the start, we dismiss any other possibility.

A further point is that a corpus view can hide patterns of individual preferences which, it is contended, should be considered if we make statements about availability of the idiom principle to language learners. A short example illustrates the problem: a search in a learner corpus may tell us that EFL learners in addition to 'correctly' using the verb DISCUSS with a direct object, use it with a preposition *about.* As such several cases of DISCUSS *about* would look like occasional mistakes leading us to conclude that learners in contrast to native speakers tend to fall into error which, according to Wray's logic, is a sign of their operation on the open-choice principle and separate storage of single words. However, it may turn out that while some learners use DISCUSS with a direct object 'in a nativelike way', others always use it with *about,* whether it has become entrenched as a result of initial L1 influence or on the basis of its similarity with a pattern TALK/SPEAK/CHAT *about*. It is also possible that they have picked it up in this form from their linguistic community, or a community of practice (Lave and Wenger 1991; see also Hynninen 2013), which can be, for example, an academic ELF community. Anyway, such evidence would speak in favor of their relying on the idiom principle rather than the other way round since the recurrence of a pattern in one learner's use, whether it is nativelike or not, points to its holistic storage, since otherwise it would not be clear why the learner would every time construct the phrase from separate words in the same way.

All these problems taken together, differences in understanding phraseology or phraseological tendency in language, bias in methodological frameworks and other contradictions, allow us to remain somewhat sceptical about the infallibility of the conclusions concerning the mechanisms of second language acquisition and use we are presented with.

3.5. An alternative explanation

It is a well attested fact that second language learners very often do not reproduce MWUs with a nativelike accuracy, and it has to be accounted for somehow. As we have seen, the most widely accepted explanation is the suggestion that second language learners tend to break down formulaic sequences into separate building blocks instead of treating them holistically when acquiring, storing or using them (Wray 2002). Therefore, according to this view, the errors in learners' use of MWUs stem from the inherent differences between first and second language processing mechanisms. However, as has been discussed above, this view is not free from problematic reasoning.

At the same time there is an alternative explanation of the phenomenon which comes from the studies of English as lingua franca (ELF). Mauranen (2011) in her article "Learners and users – who do we want corpus data from?" discusses the common ground and the points of departure between learner and ELF corpora. She points out that perhaps the main distinction between the two "boils down to language as an object of study vs. language as a means of achieving particular objectives in real environments" (Mauranen 2011: 164). Yet, despite this fundamental difference, the domain of cognitive processes appears to be an important common ground for L2 users and learners. With respect to the use of MWUs, L2 users are similar to learners in that they "tend to get them slightly wrong" (Mauranen 2011: 166). However, Mauranen prefers to call these departures from standard forms "approximations" rather than downright errors. She argues that "since they [MWUs] are useful building blocks, their approximate forms may work just as well for the purposes of facilitating communication" (Mauranen 2011: 167). And therefore, "the common use of not quite native-like phraseological units requires a use-based rather than learning-based explanation" (Mauranen 2011: 155). In other words, not quite nativelike phraseological units in learner production should not be explained in terms of the differences between L1 and L2 acquisition and processing but by the circumstances of second language use. For example, we can look at the relation between the variability of the form of the phraseological unit used and the potential amount of input received by the user. Below, I will argue that the less priming the language user has, the less exact is the form of the unit.

The term "approximation" is suggested in Mauranen 2005 where she analyses examples of discourse reflexive structures from the ELFA corpus[35] in terms of possible lexical, phraseological and functional simplification in ELF. On the whole, she contends,

[35] Corpus of English as a Lingua Franca in Academic Settings.

"ELF communication is primarily oriented to meaning rather than form" (289), therefore it is possible that "distinctions of form become salient when they are perceived to separate meanings in important ways" (290). Provided that mutual comprehensibility is achieved, Standard English forms can be ignored in ELF. Consequently, a minor often "unorthodox" variation in the form of a lexical item which does not affect the meaning expressed may be called "approximative" (Mauranen 2005: 285). In other words, approximation is "the tendency of ELF speakers to latch on to salient features of a phraseological unit, which they use in its established sense, but without exactly reproducing the standard form" (Mauranen 2009: 230).

The approximations ELF users introduce into English native language (ENL) phrases can be described as lexical and structural substitutions (Mauranen 2011). Interestingly, the finding resonates with Howarth's (1998) observations: as we remember, he has identified two main types of deviation in collocation usage in both learner and NS data, and they are grammatical modification and lexical substitution. For example, while in NS academic talk as represented by the MICASE corpus the phrase *a few words about* is a fixed cluster, ELF speakers display first, certain variability by using e.g. *couple of words abou*t as well as *a few words about*, and second, a clear preference of their own for *some words about* (Mauranen 2010). Grammatical modifications are very often embodied by non-standard use of prepositions: *discuss **about**; obsession **in**; we're dealing what is science; **on** this stage, to put the end **on** it, take closer look **to** the world, **on** the end* (from Mauranen 2011, see e.g. Mauranen 2010 for more examples).

In Mauranen 2009, it is shown that a non-standard phrase *in my point of view* used by an ELF user which at first sight may look as an idiosyncratic error or L2 idiolect, in fact, turns out to be a recurrent string in ELF communication which rejects this initial idea of a "typical" L2 learning-based error. The phrase with a further variation *in/on my point of view* is used by speakers of English from different mother tongue backgrounds, which also excludes the possibility of L1 transfer. Furthermore, it is clear that pragmatically the usage is absolutely appropriate. Therefore, the only difference to most native usage is a slight variation in the position of a preposition.

Incidentally, prepositions used in the phrase do not vary at random. All of them: *in, on, from* belong to a group of prepositions with originally "locational meanings" (Pullum and Huddleston 2002). "In formulating expressions about spatial relationships, typically one entity is taken as a reference point (or area) with respect to which another is located (Pullum and Huddleston 2002: 648). In our case *view* or "opinion" plays the role of a reference point or “landmark”, as Pullum and Huddleston call it. But even if this is a pure speculation, it is possible to hypothesise that if we look at *in/on/from my point of view* as a unit of meaning, problematic variability would seem to be a colligation of the fixed core *point of view* with a class of prepositions. As we remember, Sinclair’s unit of meaning (Sinclair 1991, 2004) integrates optional components with the obligatory core, a fixed formal component, and semantic prosody, an overall communicative meaning of a unit. While collocation is a co-occurrence of the core with a specific lexical item, colligation and semantic preference allow for structural and lexical variability, respectively. In other words, a Standard English collocation of the preposition *from* with the core *point of view* in the hands of ELF speakers loosens up and moves into the category of colligation, which is basically a grammatical generalisation, i.e. a grammatically approximated association in contrast to a verbatim association lying behind the category of collocation. In this view, colligation is a grammatical approximation of a collocation, and semantic preference is a lexical/semantic approximation of it.

Indeed, if we drop the idea of a formulaic sequence as a fixed MWU and adopt a wider approach allowing for an adaptive variability inside a unit, we will be able to see that in less frequent phrases native speakers make the same kind of approximations as non-native speakers tend to make in the phrases which are high-frequent in ENL (and which are therefore formally fixed there) but are less frequent relative to their own exposure. As an example we can take a unit of meaning with the core *naked eye* which Sinclair (2004 [1996b]) analysed himself (see a detailed discussion of this example in Section 2.2). In this unit of meaning, the core has a semantic preference for the semantic set of ‘visibility’ rather than co-occurs with a particular item and colligates with a class of propositions rather than co-occurs with a specific preposition. As such it is acceptable to say *seen with/by the naked eye*; *visible to/via the naked eye; indistinguishable* or *unnoticed to the naked eye* etc. Turning to NNS usage, it is then quite natural that certain variation appears in MWUs due to, for example, differences in frequency, which plays one of the key roles in language acquisition (Ellis 2002).

3.6. Cognitive basis of the ‘approximation’- hypothesis

In cognitive terms, such lexical and structural approximations can be explained partly with the theory of exemplar representations, partly with the long established fact that memory for meaning is stronger than memory for surface structure.

"The central idea behind exemplar-based models is that mental representations consist of memory traces of specific tokens" (Gahl and Yu 2006: 213). Therefore, each instance of linguistic experience has an impact on the system of representations. In other words, exemplar representations are extremely sensitive to the factors of recency and frequency, they "keep track of usage, allow for the representation of gradience in structures, and allow for gradual change" (Bybee 2010: 14). If we return to ELF communication, we will see how each instance of a non-standard usage, e.g. *in my point of view* instead of *from my point of view*, strengthens this representation for both the speaker and the hearer, making the probability of its further occurrence higher. At the same time, if the whole item could be accurately memorised from the first exposure, it would not be clear how variability could be introduced at all. The claim that our mind retains "memory traces of specific tokens" does not mean that each token is memorised with precise detail.

In fact, up until recently it was commonly believed that "verbatim memory is lost as soon as an utterance is understood" (Gurevich et al. 2010: 46). Gurevich et al. (2010) cite a number of studies which claim that their subjects remembered only the gist of what they heard and were not able to precisely recall the language material they were exposed to even after a short period of time. In other words, as Bock and Brewer (1974) put it, "an abstract representation of the meaning was remembered *rather than* the exact words, and ... in recall the surface structure was reconstructed from this abstract representation (Bock and Brewer 1974: 841 from Gurevich et al.: 46, their emphasis). However, in a series of recall and recognition experiments, Gurevich et al. were able to demonstrate that verbatim memory exists. They did not warn the participants of the subsequent memory testing and carefully controlled for other possible intervening factors like lexical memory effects.[36] Yet, participants in the experiments were able to both recognize and recall verbatim the specific clause witnessed with above chance probability. Even when in the fifth experiment, the participants were retelling a story with an average of a six day delay from hearing the description, they "displayed a natural tendency to reuse previously heard clauses" (Gurevich et al. 2010: 70). In particular, Gurevich et al. point out that "participants reliably used a particular structure in a particular context", for example "a passive tended to be used to describe a particular scene only when the same scene had been described with a passive" (Gurevich et al. 2010: 72). In other words, not only was the specific structure of an increment

[36] That is, remembering a particular open-class word which appeared in the clause rather than the whole clause verbatim, or in other words recognising a clause by a particular open-class word which it contains.

retained, but it was also retained in association with the specific context it had been experienced in.

The existence of verbatim memory can account for the cumulative effect of frequent exposure to inherently complex units which leads to the emergence of their holistic representation in the mind. In other words, as Bybee puts it: "While the effects of frequency are often not noted until some degree of frequency has accumulated, there is no way for frequency to matter unless even the first occurrence of an item is noted in memory. Otherwise, how would frequency accumulate? […] Thus" she continues, "the verbatim form of an experienced token must have some (possibly small) impact on cognitive representation, even if it cannot be recalled accurately afterwards" (Bybee 2010: 18).

However, although verbatim memory does not seem to disappear entirely, memory for meaning is indisputably stronger. In other words, memory for surface structure is weak, it needs to accumulate with exposure in order for an exact representation of form to be gradually built. Therefore, little exposure leads to an "incomplete" formal representation, yet what is retained is meaning. Consequently, an "incompletely", or better "inexactly" memorised unit when it needs to be used is approximated in terms of its form but within the limits of the retained meaning. The approximations made seem to be either grammatical, resembling colligations, or lexical, perhaps better to say semantic, which correspond to Sinclair's mechanism of semantic preference, as was suggested in the previous section.

Interestingly, Gurevich et al. give an example of what they did not count as verbatim recall since the produced sentence differed by more than one word from the original. The participant heard the sentence: *The kids at school were amazed* ***at*** *my* ***new*** *strength,* and what s/he produced was: *The kids at school were amazed* ***by*** *my new-****found*** *strength* (Gurevich et al. 2010: 60, emphasis in the original). As we can see the produced sentence differs from the heard one by just two "infelicities" or substitutions, which do not at all seem to be random. Indeed, if we assume that *were amazed at my new strength* is an instance of a unit, it turns out that it has been approximated in predictable places: colligation and semantic preference. All the participants in the experiment were English native speakers.

On the basis of these findings, it seems reasonable to suggest that approximations in L2 use of MWUs are systematic and can be explained by the principles on which human memory works irrespective of the language or sequence of its acquisition, i.e. whether it is a mother tongue or a foreign language for a given language speaker. Furthermore, these approximations are actually evidence in favour of the availability of the idiom principle to L2 users rather than against.

3.7. Conclusion

It is widely acknowledged that multi-word units present particular difficulties for second-language speakers (e.g. Pawley and Syder 1983; Granger 1998; papers in Schmitt 2004 and Meunier and Granger 2008). The prevailing explanation of this phenomenon comes from Wray who argues that second language learners in contrast to native speakers are predisposed to breaking formulaic sequences into separate words and storing them in such decomposed form rather than holistically (Wray 2002). In such a way Wray's model of the differences between first and second language mental lexicons is able to account for the "errors" second language speakers make in their language use.

However, at closer examination, it transpires that virtually all the studies on MWUs in second language use which come to the conclusion that phraseological patterning is a major problem for NNSs are based on comparison with native speaker language either directly (comparing NS with NNS production), or indirectly (taking the conventional sequences from NS corpora and expecting NNSs to replicate the patterns). It is argued here that even though the sequences produced by NNSs may not match NS to a hundred percent, this is not the basis to suggest that their psycholinguistic processes of acquiring the language are different. Instead, if we look at NNS patterns in the same way as we look at NS patterns, it might turn out that the processing mechanisms are remarkably similar.

For example if we analyse the 'errors' that are so notorious in second language use of MWUs, we may find out that the meanings expressed are left intact, but what is subject to certain permutations is form. Furthermore, the changes that are made to form are quite systematic in that they fall either into a category of lexical substitutions, or into a category of grammatical modifications. Therefore, it would be more correct to conceptualise these departures from Standard English forms as approximations rather than errors (Mauranen 2005, 2009, 2010, 2011 and 2012).

The second problem that is common to the studies on second language use of MWUs is the lack of agreement on what phraseology actually is. It is argued here that if we adopt a wider approach to multi-word units allowing for adaptive variability inside a unit, we will realise that approximation is a normal aspect of operation on the idiom principle and is just as typical of NS use as it is of NNS use.

In cognitive terms, lexical and structural approximations which occur in second language use can be explained with the theory of exemplar representations which are sensitive to recency and frequency of use and superiority of the memory for meaning over memory for surface structure. Incidental verbatim memory (Gurevich et al. 2010) can account for the cumulative effect of frequent exposure to inherently complex multi-word units which leads to the emergence of their holistic representation in the mind (see Bybee 2010: 18). Yet since memory for meaning is stronger than verbatim memory for structure, lexical and structural substitutions are likely to occur in less frequent structurally complex multi-word units. In Sinclair’s terms it is possible that colligations and semantic preferences are approximated collocations.

4. Data collection and research methods

This study set out to examine second language acquisition, use and psycholinguistic representation of lexis in relation to the idiom principle. In other words, its ambition is to analyse lexical items from three different perspectives: their usage patterns, their source of acquisition and their representations in the mind. This overall aim has guided the collection of data and the selection of methods and often led to unconventional solutions with respect to their combinations.

The types of data to be collected were decided on in the pilot study and the trial phases of the main study, as will be described in Section 4.1. Since three types of data were to be collected for each participant, the total number of participants had to remain small. The language backgrounds of the five participants who submitted data for the study are described in Section 4.1.1. Sections 4.1.3, 4.1.4 and 4.1.5 give detailed accounts of how each of the three types of data was collected. Section 4.2.5 is especially extensive as it not only describes the procedures of collecting word association (WA) responses adopted in this study, but also discusses word association task (WAT) as a method, presenting its various applications developed in different fields of research. The important question this section strives to shed light on is what can and cannot be said with word associations as data. It also aims to clarify what one needs to know about designing a WAT. This discussion of the method is important for the argument because the way a WAT was applied in this study is in many ways different from previous studies.

The section on WAs closes the first main part of this chapter devoted to data collection procedures. The second part, Section 4.2, focuses on the methods of analysis. The biggest challenge of this study was to find the methods with which the rich usage and word association data could be analysed in a comprehensive way. Therefore all the procedures undertaken are described in this section with a particular emphasis on how each of the methods chosen was able to contribute to the systematicity of the analysis.

4.1. Data collection: context and methods

Overall, the study was driven by an interest in L2 productive acquisition of vocabulary and the mechanisms underlying fluent and successful lexical choices. The decisions about the kinds of data to be collected for the study are based on this general interest.

4.1.1. The context and arrangements of data collection

The primary data collection was preceded by a pilot study which helped to fine-tune the envisaged data collection strategies. From the start it was decided to collect written data rather than spoken since a focus on writing allowed tracking the development of vocabulary in the writing process over time and yet maintaining the authenticity of the context, while also keeping an eye on the relevant input in a way that would not have been possible for spoken language. These criteria brought me to the Language centre of the University of Helsinki, where students in English language courses submit several written texts during a period of one semester (or half a year), which conveniently adds a time dimension to the data that can be collected.

In the course of the pilot study, I collected essays written in the framework of two courses in EAPS (English Academic and Professional Skills) from 11 students in total and analysed them with the Lexical Frequency Profile (LFP) (Laufer and Nation 1995). The LFP, which measures the lexical richness of writing, was generated for each writing sample, and since each student submitted more than one essay, it was possible to see the development of profiles over time. However, it became clear that although the LFP measures seemed to be able to track the development of lexical richness, they could not give any insight into how vocabulary was actually acquired and what exactly changed in the lexical quality of learner writing with the changes in the profile. At the same time it proved to be promising to analyse lexical strings taken from the essays with Sinclair's categories of co-selection: core, collocation, colligation, semantic preference and semantic prosody, discussed in Chapter 2. It was interesting to notice that all the five categories were distinguishable in learner writing, which tentatively suggested that learners operated on the idiom principle. In sum, the pilot study showed that in order to make any conclusions about the processes underlying lexical production, including the idiom principle, other, richer data was needed from the same writers, which also meant that the number of participants had to be smaller.

Given the insights from the pilot study, it was decided that the most suitable ESL writers for the study were the students writing their Master's theses in English, a non-native language for them, which is a relatively common phenomenon at the University of Helsinki. Already at the graduate level students are acutely aware of their potential audience and often do not want to restrict themselves to the national audiences only. For example, one of the students I worked with said she was writing for her Swedish colleagues over and above Finnish. It is important to point out that the data was thus collected in a typical academic ELF environment: the writer and the audience did not share a common native language and therefore selected English as their tool for communication

Collecting drafts of Master's theses gives several advantages over collecting essays, or texts also often called compositions, written in the context of formal instruction in English. Master's thesis is an established genre of academic discourse with a clear "set of communicative purposes" (Swales: 1990: 46), and in that it can undeniably be regarded as naturally occurring data. In contrast, an EFL composition serves as a pedagogical tool which enables practice and evaluation of English language skills. It can be argued that it forms a genre of its own but a genre which has less relevance outside the TESOL community. Therefore, it is difficult to find comparable material as a point of reference for EFL essays. In contrast, Master's theses have a very clear reference point: being written in a certain disciplinary discourse community, they need to follow its disciplinary practices, and therefore can be compared to other academic texts which belong to the same field. Also, theses are written over a substantial period of time and thus are suitable for a longitudinal study.

The drafts of Master's theses and other data in this study come from students I met at the Language Centre. Preparing to write their theses in English, students often take the course "Advanced EAPS: Academic Writing for Study Purposes". The course is intended for students who are writing academic texts in English and consists of 14 classes devoted to different aspects of academic writing from the question of cohesion to the problems in punctuation. During the course the students write a one-page summary and a critique of the same length preferably related to the topic of their Master's thesis. There is an additional possibility of signing up for two individual consultations with the teacher of the course. For each consultation a student submits a 5-page draft, for example a draft of a thesis chapter, and receives language feedback from the teacher at the consultation: the framework I later adopted for my meetings with the students. I followed the course in the spring semester of 2009 to familiarise myself with the contents of the course and find student volunteers for the study.

The work with the first two students (of the total of seven) was very much experimental. It was clear that the consecutive drafts of different parts of their Master's theses adequately met my criteria for the usage data. However, I was not sure what kind of data can complement this usage data to gain additional insight into the mechanisms underlying their lexical choices. We arranged one to one and a half hour meetings each time the students had a five- to fifteen-page draft ready. I used these meetings to interview the students and experiment with some lexical tasks often used in Applied Linguistics. In this way I tried out a Word association task (e.g. Fitzpatrick 2006, 2007), a Vocabulary Levels Test (Nation 2001), a Vocabulary phrase gap-filling task (adapted from Schmitt et al. 2004). I also created my own version of a gap-filling task where different components of extended units of meaning (collocation, colligation, semantic preference or even the core) were deleted and had to be supplied. All the meetings were recorded, and the students' consent forms to use their samples of writing and recordings of the meetings were collected.

Experimenting showed that the most promising approach was to test the words the students used in their writing by employing them as stimuli in a word association task (WAT), so that each student had his/her own set of stimulus words. In this way it was possible to compile "word profiles" (individually) for each writer, consisting of his/her word associations, concordances of recurring patterns retrieved from his/her writing samples and comments on his/her word associations and lexical choices recorded during the individual meetings. This is how it was decided that word association responses would comprise the second major type of data. Experimentation also helped decide on the types of words to be selected as stimuli for the WATs. I will give more details about the rationale for and the administration of WATs in Section 4.1.5. Later it became clear that the "word profiles" could benefit greatly from comparison to the language served as a source of acquisition for the observable patterns. The collection of this type of data is discussed in Section 4.1.4. Section 4.1.3 specifies the procedures in the collection of usage data, while Section 4.1.2 provides more information about the language backgrounds of the participants first.

4.1.2. Participants

Thanks to the collaboration with the Language Centre, I eventually had seven participants (including the two mentioned above), with whom I worked individually and collected their thesis drafts longitudinally over a period of about a year on average, depending on how fast they were able to progress with their studies. The students were motivated to continue working with me since I offered feedback on the language of their drafts in exchange for their help with data collection. Out of seven complete data sets, it was later decided to exclude two from the analysis to sharpen the focus. The first of these data sets belonged to the student whose word associations were collected at the beginning of the study when the methodology of selecting stimuli for the WATs was not yet fully developed. The second one was left aside due to the student's biographical background which markedly stood out from a more or less homogeneous group the other students formed.

The homogeneity of the final group of five is contained in many language-related aspects of their biographies.[37] Four of the students, Kaisa, Linda, Hertta and Maisa, speak Finnish as their mother tongue, and one, Nora, Swiss German.[38] That is, for all of them English is a foreign language. All the students were born between 1981 and 1985 belonging to the same generation and therefore having been exposed to globalization and the spread of English used as an international language or a lingua franca to a similar extent. All of them had between7 and 12 years of formal English instruction at school and 2-3 courses at University, except for Kaisa who studied in a Finnish-English bilingual school and obtained her Bachelor's degree in Germany. None of them had lived in an English-speaking time for a considerable amount of time, their stays ranging from 6 weeks to one year in total. Since Nora is the only one whose mother tongue is other than Finnish, she is also probably the one using English in her everyday life in Finland most of all as she studies in English at university and uses English in addition to German with her friends. Still, Linda and Kaisa also mention English as a language of their everyday communication, Hertta in contrast mentions Swedish. All of the students are multilingual, listing three to four languages in their repertoire, in addition to the mother tongue and English. The participants represent different fields: Nora and Linda study Communications at the Faculty of Social sciences, Maisa is a demographer, Kaisa a computational linguist and Hertta's field is Archaeology. This diversity of academic disciplines is advantageous for this study and was deliberately sought for: while being sufficiently homogeneous as a group enabling intersubject comparisons, the students do not come from exactly the same setting which raises the generalisability of the findings.

Another reason why the group of participants needs to be homogeneous stems from its small size: in case the data analysis shows that, bearing the research questions in mind, there are important differences in the participants' acquisition, usage and processing of units of meaning, there will be no possibility to examine any additional variables and say why the differences have arisen. In that sense, it is somewhat risky to have a linguist as a participant since her attitude to language and perceptions of the ways one needs to handle it might be different from other students leading to divergent results. Yet, Kaisa studies computational linguistics which, judging by her drafts, is a more technical field than traditional philology.

From a more general perspective, the selection of participants might not seem ideal, and yet, it is maintained, it serves the purposes of the present study well enough. For

[37] The biographical details of the students summarised below were collected during the interviews and with the help of questionnaire designed for the purpose.

[38] For ethical reasons, the real names of the students are not disclosed, instead all of them are given pseudonyms.

example, from a sociolinguistic point of view, it might be desirable to have participants of different gender. However, it would be odd to hypothesise that male and female language users cognitively process language differently. Or, in traditional SLA studies researchers prefer to have participants from different language backgrounds since the assumption is that mother tongue has a very strong influence on any language acquired subsequently. This is not the view adhered to in this study and in any case there is no intention to investigate language transfer. As for the number of participants taken for the study, it had to strike a balance between enabling certain generalisation of findings and requiring a reasonable amount of time for data analysis. The findings yielded from the analysis of fewer than five data sets may appear questionable. At the same time it will become clear form the description of the types of data collected from the participants which follows below that it would not be feasible to analyse data from more than five participants in terms of the time it would require.

4.1.3. Longitudinal corpora of written production (C1)

As a result of the work with the students, described in Section 4.1.1, three types of data were collected for each of them forming individual data sets: (1) a corpus of written production, (2) a reference corpus of expert writing in their field and (3) word association responses together with the interview data providing their retrospective comments on the associations they have come up with. I will start with describing the compilation of the corpora of written production.

The longitudinal corpora of written production (C1) are compiled from students' MA thesis drafts submitted consecutively and written over the period of time when the students were working on their theses. Kaisa's C1 also includes the summary and the critique written during the course she attended described in Section 4.1.1. The writing samples in the corpora are unedited and therefore represent students' own usage. Table 4.1 below gives the lists of drafts collected from each student with the dates of their submission or time periods when they were written, to provide at least a rough view of how longitudinal the data collection was and what the time intervals between the draft submissions were. The last row also gives the respective sizes of each corpus in word tokens.[39]

Table 4.1 The directory of the drafts collected and the sizes of resulting corpora

Kaisa	Hertta	Maisa	Linda	Nora
kaisa_1_170309.txt	hertta_1_spring2010.txt	maisa_1_151009.txt	linda_1_150410.txt	nora_1_060510.txt
kaisa_2_140509.txt	hertta_2_spring2010.txt	maisa_2_191109.txt	linda_2_290410.txt	nora_2_201010.txt
kaisa_3_250609.txt	hertta_3_spring2010.txt	maisa_3_140110.txt	linda_3_030510.txt	

[39] Word tokens were counted using ConcGram (Greaves 2009).

kaisa_4_100809.txt kaisa_5_140909.txt kaisa_6_300909.txt kaisa_7_021209.txt kaisa_8_130410.txt kaisa_9_300410.txt kaisa_10_300410.txt kaisa_critique_110209.txt kaisa_summary_210109.txt	hertta_4_0510.txt hertta_5_130910.txt hertta_6_1010.txt hertta_7_050111.txt hertta_8_260111.txt	maisa_4_140110.txt maisa_5_031110.txt maisa_6_07-281210.txt maisa_7_07-280211.txt maisa_8_010311.txt	linda_4_040511.txt	
21,887	39,449	21,501	37,641	14,843

As evident from the table, the resulting corpora are small, different from each other in size and number of drafts comprising them. However, these limitations are natural for this type of data, individual written production, and since the corpora are not going to be compared to each other, they do not hinder the analysis to be carried out in any serious way.

4.1.4. Reference corpora of the priming language (C2)

The individual corpora of written production allow investigating lexical choices each student makes. Looking at these lexical choices in its turn raises a question of where they come from. A usage-based view of language predicts that language is acquired through experience with input, and thus it should be possible to track down the students' lexical choices to their experience of the language: "language learning is estimation from sample" (Ellis 2009: 139). Therefore it was decided to compile individual reference corpora of the priming language (C2) for these students consisting of the texts or sorts of text they were likely to read when preparing for writing their theses, i.e. texts from the same field as their own written production. I call these corpora corpora of language exposure, source of acquisition corpora, corpora of expert writers or simply reference corpora.

Defined this way, C2 corpora can and in fact do include texts written by non-native speakers of English, who nowadays outnumber native speakers in general and in academia in particular (e.g. Graddol 2006). This is not in conflict with the purposes of the study since I am not interested in the extent to which L2 users can acquire native-like lexical patterns, but the extent to which they can acquire lexical patterns from the input, which is in line with a usage-based understanding of language acquisition. Therefore, L2 users' texts are going to be compared to the language these users were exposed to, rather than native-speaker data.

Importantly, it is not argued here that academic texts the student read were their only priming language. It is obvious that they were exposed to all kinds of other language, and therefore general purpose corpora like the BNC are also relevant as possible points of comparison. However, it is not feasible to compare a corpus of somebody's thesis drafts with the BNC on a systematic basis: while there certainly will be some overlap, they are not directly comparable due to the register-specificity of language (see e.g. Biber 1988; Biber et al. 1999). Therefore it was decided to compile corpora which would be representative of both the priming language of the students and the kind of language use which the students would most probably target at in writing their theses. (See Section 4.2.5 for more discussion of the use of the BNC.)

The texts for C2 corpora were selected on the basis of the reference lists the students provided in their theses and the number of times each work on the reference list was cited in the body of text. Effort was made to find those publications which were cited two or more times in the thesis to ensure that that they were not only cursorily referred to but indeed were likely to form an important part of the student's background reading.[40] Table 4.2 below provides information on the size of each of the individual priming corpora (C2) together with the size of the corresponding written production corpora (C1) for comparison. The last row shows how many times larger the priming corpora are than the written production corpora.

Table 4.2 The sizes of individual written production corpora and individual priming corpora

Student	Kaisa	Hertta	Maisa	Linda	Nora
C1	21,887	39,449	21,501	37,641	14,843
C2	64,809	123,662	85,234	133,778	86,500
C2/C1[41]	3.0	3.1	4.0	3.6	5.8

It was considered important that the priming corpora are as large as possible since they had to give a good picture of the kind of language the students were likely to be exposed to. Unquestionably, the students were exposed to much more language than is contained in the reference corpora. However, if the corpora are representative of this exposure language, comparisons between C1 and C2 should be feasible.

In contrast to C1 corpora, C2 corpora include texts of different formats and genres. Clearly, students' academic priming consists of different kinds of texts even if we are looking at field-specific priming only, the kind that they are expected to mostly fall back on in writing their theses. And indeed the students used not only academic articles published in journals, edited volumes and conference proceedings, but also magazine articles (like Linda), theses (like Hertta) and WHO reports (like Maisa). Table 4.3 presents an overview of each student's C2 corpus.

[40] A World Health Organisation (WHO) report and a U.S. Census Bureau report used for Maisa's reference corpus (see Table 4.3) form an exception: these are not the actual reports cited but are similar to them and produced by the same organisations. The two reports were included in C2 instead of the original ones for reasons of availability.

[41] The size of C2 divided by the size of C1.

Table 4.3 The composition of students' C2 corpora

Student	Kaisa	Hertta	Maisa	Linda	Nora
C2 files	5 journal articles 7 proceedings papers 1 edited volume paper	3 journal articles 1 MA thesis 1 introduction part of an article based PhD 7 edited volume papers[42]	3 journal articles 5 reports (2 Ministry reports, a WHO report, WHO guidelines, a U.S. census bureau report) 1 population projection software guide	19 journal articles 1 BBC press release 2 magazine articles	10 journal articles 1 edited volume paper
C2 size	64,809	123,662	85,234	133,778	86,500

It is assumed that all these types of text will be part of the students' priming underlying the lexical choices they make in their theses. At the same time, it is important to bear in mind that C1 and C2 corpora are not comparable in everything that concerns genre: certain genre-specific patterns may be different (see Section 5.4.2).

As for the processing of corpus files, I tried to retain as many meaningful stretches of text as possible and therefore included parts of text other than its main body, like abstracts, acknowledgements, footnotes, endnotes, tables, figures, their captions and bibliographical references, when there was a reason to do so. Most of the texts were converted from PDF to Plain Text using PDFX software (PDFX v1.8). At the stage of clean-up, all the corpus files were manually proofread to make sure that automatic text conversion did not result in any unusual symbols or broken down chunks of text. But the focus of attention mainly lay at phrase-level, matters like missing paragraph breaks were not taken into account.

4.1.5. Psycholinguistic data: word association responses

As mentioned in Section 4.1.1, word associations were chosen as a complementary type of data from among other possible alternatives. A word association task (WAT) gives an insight into the psycholinguistic aspects of production for any set of target words, which was exactly what was needed in this study. Having data on how the words were used, the study called for a different perspective on the same words: knowing what lexical choices are made in use, the

[42] Out of seven edited volume papers, six come from the same two volumes due to availability, however since all these papers were cited in the student's thesis, it seems that she herself relied on these two books extensively.

question arose how they might be represented in the mind. Vocabulary Levels Test would only be able to give an indication of the participant's overall size of receptive and productive vocabularies, gap-filling tasks would illuminate some aspects of the productive knowledge of the words tested, lexical priming tests, like a lexical decision task, do not require the subjects to produce words and thus probe receptive rather than productive skills, in addition to having a more complicated set up than word association tasks. Word association tasks certainly have their limitations, which will be discussed later in this section, but they give an opportunity to test large numbers of target words in very little time and examine which words are produced (rather than merely recognised) in response.

In what follows, I first discuss the different applications of the word association method in research over its history of more than a century. The use of the WAT as a method has been somewhat controversial, so it seems essential to understand what can and cannot be said with WAs as data and how a WAT should be designed. After this brief review, I go back to the present study and describe the procedures carried out to collect word association responses from the students in such a way that they would complement and inform the analysis of the usage data. Limitations of the task and its implementation in this study are pointed out at the end of the section.

Word association task as a research method

An interest in a word association task, originating from the studies in psychology and psychiatry dating back to be the beginning of the 20th century, has persisted until now despite the rather varying degrees of success with which it was applied. The basis for this consistent interest is effectively summarised by the compliers of the Edinburgh Word Association Thesaurus: "…it has always been and remains to be a general belief that associative processes are a basic component of thought and cognitive processes in general. It has been the hope of many investigators that the regularities of associative responses will yield some insight into the structure of the human mind" (Kiss et al. 1973: 153).

Word associations present (what seems to be) rich data obtainable at low cost. They also seem to be able to fire up the imagination of very different researchers. Over more than a century of the test being in active use, its fields of application range from psychoanalysis to SLA and from generative semantics to computational modelling of semantic lexicons. The interpretations of word associations are so diverse that they can seemingly be used to back up almost any theory. WAs have been used to distinguish between normal and insane people, native speakers and non-native speakers, to claim that language processing is rule-based and to argue for the applicability of graph theory for modelling our lexical knowledge, depending on what the theoretical presuppositions the researchers nurtured from the start were. In fact, some went as far as to argue that in order to produce the association *woman* in response to *man,* the subject would go to the list of features for *man*: [+Noun, +Det-, +Count, …+Male], apply the rule "change the sign of the last feature", calculate that it means '–Male' and thus produce the word *woman* (see Clark 1970: 274). Others find it evident that an L2 learner becomes more proficient in a language when the route between any two nodes-words in his mental lexicon becomes shorter through overall growth of the lexicon and increase in its connectedness (see e.g. Meara 2009). With so many interpretations accumulated over the century of contrasting research, it is not easy to strip away the underlying theoretical assumptions. Therefore it seems useful to look back at what different researchers were able to see in WA responses from a critical viewpoint.

Early word association research

The history of the word association test as a research method is usually traced back to Sir Francis Galton (1879) who studied his own WA responses produced on four separate occasions on the same set of stimulus words and reported his findings in the *Brain*. What impressed him in his associations (according to Levelt 2013: 148) is that 42% of them occurred more than once suggesting that they were all stored and just reproduced by memory over and over again. He found himself under the impression that word associations "lay bare the foundations of man's thoughts with curious distinctness, and exhibit his mental anatomy with […] vividness and truth" (Galton 1987: 162 as cited by Levelt 2013: 148).

The idea was shortly picked up by other researchers, and a whole number of interesting observations were made. These observations can be of interest for us now since they were not yet born out of any prevailing paradigm prescribing how WAs need to be looked at, as was often the case later. Among them is the discovery of asymmetry in WAs: temporal associations are made in chronological order rather than in the backward direction, e.g. *March* → *April* rather than *April* → *March* (Trautscholdt 1883 as reviewed in Levelt 2013: 149). Thumb and Marbe (1901), a linguist and a psycholinguist, used the WAT to test Hermann Paul's hypothesis that analogy is a psychological mechanism of language production, i.e. by analogy to previous encounters, that also leads to language change. To illustrate, *hide-hid-hidden* could in principle change to *hide-hode*-hidden* in analogy to *ride-rode-ridden* because of the rhyme association *hide-ride*. This does not happen because *hid* is a frequent and well-entrenched word, but for example children would be more prone to make the change and ‘regularise’ the verb forms because of the weaker memory traces (Levelt 2013: 152).[43] Thumb and Marbe saw a psycholinguistic support for analogical language change in the fact that word associations too tend to be made within one word class (nouns elicit nouns etc.). One finding considered to be especially important is the so-called Marbe’s law which postulates that when a stimulus word elicits the same response from different test-takers, the response time is shorter, and in fact the larger the agreement between test-takers on the response, the faster this response is (Thumb and Marbe 1901 as summarised by Levelt 2013: 152-154). They further argued that strongly associated pairs (1) become more similar in form and (2) occur in slips of the tongue.

Later in the psychological studies of the 20th century, there was an attempt to standardise word association procedures and produce word association norms which could be used for diagnosing abnormality. In a well-known study, Kent and Rosanoff (1910) collected a huge amount of data recording responses to 100 stimulus words from 1000 normal adults. After some time the usefulness of this kind of norming data for psychological purposes was questioned, but since it is rarely possible to collect that much data for smaller-scale research, Kent-Rosanoff’s database of responses continued to be used in other types of studies. This meant of course that the original Kent-Rosanoff list of stimulus words, as it is often referred to now, also had to be adopted.[44] In this way we have traces of psychological thinking even in much more recent studies in other fields, for example SLA,[45] as will be pointed out below. By now there are other databases available, for example, we also have Palermo and Jenkins’ (1964) word association norms, the Postman–Keppel lists (1970) comprised of several works on WA, Birkbeck word association norms (Moss and Older 1996) and the Edinburgh Associative Thesaurus (EAT) which includes data on 8400 stimuli (Kiss et al. 1973): it is also available online and most often used in current research.

In addition to being used as norming data, word association stimulus-response pairs strongly suggest a notion that a language user’s mental lexicon resembles a network with its nodes and links representing words and connections between them. This idea seemed to be exercised as early as 1965 by Deese, with other researchers soon proposing an application of

[43] Here we can already see the seeds of usage-based theories, exemplar representation and, incidentally, an explanation of why ELF users of today, who just like children have weaker memory representations, are capable of inducing accelerated language change.

[44] The major problem that some SLA researchers see in using Kent and Rosanoff’s list is that it consists of frequent words and therefore elicits predictable responses – an observation suggested by Meara. It is also noted that it was not intended to study linguistic behaviour (Fitzpatrick 2013, citing Meara 1983; Schmitt 1998; Wolter 2002; Fitzpatrick 2006). There are other problems with using this or a similar list for linguistic purposes as I will try to point out in connection with the discussion of Carl Jung’s method.

[45] A study as recent as Albrechtsen et al. 2008 takes their stimuli from the Kent-Rosanoff 1910 list.

graph theory to word associations (Pollio 1966; Kiss 1968; Kiss et al. 1973). It must be noted, though, that Kiss et al. make a reservation that the “associative organisation of the subjective lexicon is just one of the facets of this lexicon with others needing different methods of study” (154): a point which is not always shared or remembered.

Word association method in SLA

It is perhaps not surprising that at some point the word association method was adopted in SLA research: after all it seems intuitive that word associations responses have something to do with how we store our vocabulary. Both ideas, word association responses as norming data and as links reflecting the network-like structure of the mental lexicon, found their application here.

The idea of the network organisation of the mental lexicon is strongly supported and followed by Meara. For example, he suggested (1997) that instead of seeing vocabulary knowledge as having two dimensions, breadth and depth, where breadth is the number of words known and depth is the quality of the knowledge about each of the words known, it can be depicted as a network where the idea of breadth corresponds to the size of the network or the number of nodes in it, and depth is the connectedness of this network. The advantage of this way of seeing vocabulary knowledge is that it makes explicit the intuitively plausible assumption that the more words we know the better we know them. The word associations then should be able to shed light on the questions such as: “what does a learner's mental lexicon look like, and how is it different from the mental lexicon of a monolingual native speaker?” (Meara 2009 [1983]: 21).

While some of the SLA researchers using the word association method share the idea of conceptualising vocabulary knowledge as a network-like mental lexicon and others do not explicitly take sides, they mostly apply word association data to studying the following questions:

- What is the difference between L1 and L2 word association responses? What does this difference say about the organisation of their mental lexicons?
- Can the WAT be used as a measure of L2 proficiency? Or in other words, is there a relationship between word association responses or scores gained on a WAT and language proficiency?
- Will L2 learners’ word association responses and response patterns become more native-like with their proficiency developing?

Different studies produce rather contradictory answers to these questions. But, as we will soon see, some research shows that a native-like pattern or a pattern which can be meaningfully called canonical simply does not seem to exist. Therefore, the questions might be in fact irrelevant.

In a nutshell, the logic behind the research questions asked in SLA seems to be the following: word association responses must somehow reflect the organisation of the mental lexicon of the respondent, therefore by comparing word associations received from different respondent groups it is possible to compare their mental lexicons. In SLA it has always been of interest how L2 mental lexicon, which is seen as being at the developmental stage, differs from a mental lexicon of a native-speaker, which is viewed as a target for non-native speakers. Having adopted word association methodology, SLA scholars then try to (1) classify word association responses (e.g. syntagmatic/ paradigmatic/ clang) and (2) rate or compare L2 individual responses or classified profiles with those of either more proficient non-native speakers or native speakers (which are by default more proficient). With the rise of the level of proficiency, L2 vocabulary knowledge is supposed to be developing in its breadth and depth, which must have an impact on the organisation of the mental lexicon. Therefore, researchers usually expect shifts in association responses, for example from high-frequent to low-frequent or from more idiosyncratic to canonical, i.e. responses frequently given by native speakers (Albrechtsen et al. 2008). Also, shifts in the type of responses given are expected, for example from clang or form-based to syntagmatic and paradigmatic or as more recently has been suggested, from paradigmatic back to syntagmatic at a more advanced level (see Fitzpatrick 2006 or 2007; Albrechtsen et al. 2008).[46]

The rationale behind the prevailing comparison of L2 word association results with native speaker performance rests on two interrelated assumptions, namely that native speakers have a well-developed and fully functional mental lexicon with all the necessary links in it and that native speakers are homogeneous as a group in the way their mental lexicon is structured. However, Fitzpatrick (2007) showed that adult native speakers varied considerably in their response behaviour, and whereas one showed a preference for collocational associations, another could be keen on giving defining synonyms in response to word stimuli. The plausibility of distinguishing between native-like and non-native-like or canonical and non-canonical responses also seems questionable. This methodology is adopted by Albrechtsen et al. (2008), even though they explicitly say that "the variation in the

[46] In all, word association research in SLA has not been able to produce conclusive results on the direction of the shift if there is one.

different lexical items supplied by native speakers as responses to a stimulus word is extremely high" (34), and that "some of the stimulus words gave rise to so many different associations that it was decided that only responses given by roughly ten per cent or more of the informants from the norming groups could meaningfully be classified as canonical links" (46-47) which means that 90% of NSs give non-canonical responses, i.e. it is quite canonical to give a non-canonical response. To solve this problem Schmitt (1998b) suggests that a response a NNS gives can be rated along the continuum of typicality and proposes a specific method of how this can be done. However, adoption of this approach would still mean that since NSs give a whole array of diverse responses, they themselves can be more and less 'native-like'. To sum up, native speakers are not homogenous as a group in either giving normative responses or favouring a particular type of responses.

It is interesting to note in this respect that apparently Kent and Rosanoff did not see anything wrong in including NNS word association responses in the norming data since this is what they actually did, describing their respondents in the following way:

> Many were from Ireland, and some of these had but recently arrived in this country; others were from different parts of Europe but all were able to speak English with at least fair fluency. (Kent and Rosanoff 1910: 38-39)

Fitzpatrick (2007) also makes a remark about this interesting fact. In the same article, she unpacks many other commonly held assumptions about word associations: that NSs are homogeneous as a group in producing similar word association responses or similar types of them or that it is children who give predominately syntagmatic responses meaning that paradigmatic responses are somehow 'more advanced'. However, I feel that the fundamental questions of what word associations can and cannot say and what it is that they are probing into have still not been addressed. To at least faintly illuminate these questions, I would like to go back to one of the early applications of the word association method which is also very well-known: Carl Jung is reported to be the first to use the method clinically in the beginning of the 20th century. Since the word association test is originally adopted from the field of psychology, it is thought that it might be useful to look at least at one example of its application in this field to get a grasp of the underlying assumptions with which it was used there. One point which seems important about Jung's experiments is that he seemed to be content with the method, while for many other researchers the method produced more questions than answers.

Jung's application of the word association method

I will base my understanding of Jung's approach on his article which appeared in *the American Journal of Psychology* in 1910 and consisted of the three lectures he gave on the topic, the first of which directly focused on the association method, his application of it and the way in which it was useful to him in his work. It seems to me there are three distinctions which characterise Jung's approach to the association method in a crucial way: (1) his treating association responses as reactions to real situations rather than words, (2) his focus on the typology of the responses rather than the responses themselves, and (3) his interest in the general performance of the test person on the task, e.g. his/her ability to follow the instructions and the response times. We will now look at each of these three distinctions in more detail.

In relation to the first distinction it can be said that, in contrast to WA studies in Linguistics and SLA, Jung was not interested in the associative links of a word. His interest was in the referent designated by a word. This can already be clearly seen from the types of words he chose for his experiments: these are words which have very clear referents in the real world, like *green*, *head*, *to pay*. To quote Jung, he used words as "linguistic substitutes of reality", expected that "the stimulus word will as a rule always conjure up its corresponding situation" and was interested in the reaction of his patient to this situation rather than the word itself (Jung 1910: 224-225).[47] A WAT was also in a way a conversation starter for him. For example it was not always easy for a young lady in the beginning of the 20th century to talk about her feelings towards marriage, but her response to a stimulus word 'bride' or 'bridegroom' could give a hint to the doctor whether this is a topic worth of further attention or not.

It seems that the patients treated stimulus words in their task in exactly this expected way: they were indeed searching for the idea, the concept behind the stimulus word to which they needed to react. It can be seen from the following examples:

to quarrel - angry - different things - I always quarrel at home;
plum - to eat - to pluck - what do you mean by it? - is it symbolic?
to sin - this idea is quite strange to me, I do not recognize it
(Jung 1910: 228-229)

[47] "If I were a magician I should cause the situation corresponding to the stimulus word to appear in reality and placing the test person in its midst, I should then study his manner of reaction. The result of my stimulus words would thus undoubtedly approach infinitely nearer perfection. But as we are not magicians we must be contented with the linguistic substitutes for reality; at the same time we must not forget that the stimulus word will as a rule always conjure up its corresponding situation" (Jung 1910: 224-225).

There are two main reasons which can account for this fact. The first is the context of situation: the test persons were patients who came to see a doctor about their psychological problems. So it should not be surprising that by virtue of this context, the test persons were absorbed in their own psychoanalysis and indeed tended to treat the stimulus words as cues to their inner feelings and emotions. Since a word association test is a very loose task, it is open to interpretation and can be approached in a number of different ways. An approach chosen also becomes a strategy of giving responses. The second reason is the combination of stimulus words selected for the task. As we have already seen, Jung was interested in particular real life situations and hoped that the stimulus words would evoke them, so he chose the stimuli accordingly: all the words are easily imaginable. Reinforcing the effect, all the words are given in their base forms which encourages thinking in terms of meanings and concepts rather than their linguistic realisation, making syntagmatic responses quite unlikely. And when stimulus words in a test are arranged so that words of a similar type follow each other, it is easy for a test-taker to lapse into a certain pattern of thinking and producing responses. WA studies in Linguistics and SLA select similar types of words for stimuli and present them in a very similar way, with many of them simply adopting Kent-Rosanoff's list,[48] while their purpose is completely different: to probe the language knowledge of the test-takers rather than their psychology.

The second characteristic feature of Jung's experiments is the focus on the patterning of the responses rather than the responses themselves. He noticed, for instance, that patients who are lacking in intelligence, know about it, find it painful and as a result are diagnosed with an intelligence-complex, take the task as an intelligence test and respond with definitions of the kind: "table, - a piece of household furniture" or "to promenade, - an activity" (Jung 1910: 236). Likewise, test persons suffering from emotional deficiency respond with associations of the predicate type which are also strongly emotional, like: "piano – horrible" or "mother - ardently loved" (Jung 1910: 237). In other words, it seems like while the test persons of the first type ask themselves: "What do I know about these words?", the test persons of the second type are answering the question: "What do I feel towards these things, ideas, concepts?" So one observation to be drawn from this is that in certain context the interpretation with which test takers approach the task might be very

[48] The notable exception is the EAT, where functional words and different word forms are included as stimuli even though this is a side product of the main goal rather than conscious decision on the part of the researchers, as it seems. The authors collected WAs with the aim of building semantic networks, so they used the responses of the first set of respondents as stimuli for the second, the second for the third etc. For this reason such words as *if* and *finds* rather than simply FIND crept in purely by virtue of being somebody's responses.

important and constitute one of the results of the test. But is also important to note than Jung did not attempt to collect a database of word association norms or any other type of ‘canonical’ responses, which is one clear trend in later applications of the word association task.

And lastly, as a psychologist Jung was particularly concerned with different aspects of the test-taker’s performance on the test as they enabled him to make conclusions about the person’s character and diagnose his/her problems. First, for Jung WAT was a handy simple enough task to let him an opportunity to observe his patient in action, how s/he is able to perform on a task, just like any real life task: Is s/he able to follow the instruction and, so to speak, play the game? Is s/he nervous in completing the test? Does s/he manage with the test? Second, for him the test-taker’s reaction times to different stimuli were crucial for the analysis and diagnosis. Longer reaction times indicated possible barriers the patient had to overcome in order to give a response: so the stimuli producing them would constitute potential points of interest for further enquiry. And third, in addition to a word association task itself, Jung also used a follow-up “reproduction test” where he asked the test persons to reproduce the associations they had just responded with. In those cases where memory failed, Jung suspected “an emotionally accentuated complex” (Jung 1910: 238) since while experiences which are highly emotional are indeed memorable, as is commonly known, the linguistic realisation of them is not (in simple terms, we do not remember the exact words which accompanied an emotionally strong experience).

What Jung’s example seems to be showing is that such factors as the context in which the WAT is taken, the interpretation of the task by the test-taker, the selection of stimuli and their arrangement can completely change the nature of the task and the responses elicited. Perhaps this could be one of the reasons why different applications of the WA method yielded such contradictory results. It seems that the test needs to be carefully designed, and the context needs to be taken into account.

Combining corpus linguistic and WA data

A relatively new trend in methodological solutions is to compare human word associations to associations found in corpora. For example, Mollin (2009) compares word associations from the EAT database with corpus collocates from the BNC in order to find out whether the two kinds of data can be combined to provide an insight in to the phenomenon of co-occurrence.

In particular, she compares top ten WAs in EAT and top ten collocates in the BNC as measured by their absolute joint frequency and four different statistical tests (MI, z-score, MI3 and log-likelihood). In neither of the cases does she find any overlap between the WA and corpus lists: absolute frequency, MI3 and log-likelihood favour grammatical collocations such as *of* → *the*, while MI, z-score have a low-frequency bias and retrieve combinations like *chugga* → *chugga*, which simply do not occur outside the company of each other. But there are no such stimulus-response pairs in EAT, its top three being *lob* → *ear, cheddar* → *cheese* and *hong* → *kong*. Mollin also tries out another approach: randomly selects 30 words used as stimuli in EAT and compiles all the responses they elicited as well as all their collocates (frequency > 5) from the BNC together with their collocation strength values obtaining a list of 20,003 word pairs in total. She finds that the frequency of the response elicited in WAT does not predict collocation strength on any of the five measures and vice versa. Likewise, there is no correlation found between frequencies of corpus co-occurrences and EAT values when only those word pairs are compared which occur in both the BNC and EAT.

Even when WA responses and BNC collocates of the same stimulus/node words are compared qualitatively, differences are apparent: some of the WA responses are syntagmatic and correspond to the BNC patterns, like *afraid* → *of* or *time* → *machine*, but the majority of the responses are semantically related, like *afraid* → *terror* or *frightened*, which expectedly do not frequently co-occur. Some stimulus words, like *time*, seem to elicit more syntagmatic responses which also match corpus co-occurrences. But, overall, the comparison suggests that there is still too little overlap to talk about any kind of systematic similarity.

In other words, as Mollin concludes, even though it is advisable to find a way to complement corpus observations with psycholinguistic evidence, word associations cannot fulfil this function. Mollin agrees with Clark (1970) that WAT is a semantic task in which the mental lexicon is searched for related words, and therefore, it has little to do with the language production mechanisms (Mollin 2009: 196-197). That is, it is suggested, we are able to produce word associations because we can understand and produce language and not the other way round: word associations do not lie behind our mechanisms of language comprehension and production (Clark 1970: 272).

However, it seems there are certain aspects in Mollin's research design which could have predetermined the lack of correlation found. First of all, the EAT and the BNC are not necessarily directly comparable. The subjects in EAT are all 17 to 22 year-old mainly undergraduate students from British universities whose WA responses were collected in the early 1970s at their study places, while the BNC is a general corpus of British English completed in early 1990s and comprised of a diversity of texts from different language domains. So there is a difference in time and settings of data collection as well as in age and social class of informants. In addition to this, the size of EAT is 8,400 stimuli multiplied by 100 responses collected for each stimulus, that is 840,000 WA responses, the size of the BNC is 100 million words, which makes it ca. 120 times larger. The BNC texts have 3,294 authors behind them (see http://www.natcorp.ox.ac.uk/docs/URG/ for information about the design of the corpus), while the EAT associations come from 8,400 subjects. If two corpora with such parameters were compared and found dissimilar in language patterning, this would not be taken as a surprising finding.

There are two more points that can be made. First, the statistical measures used to retrieve collocates from the BNC to compare with the most common stimulus-response pairs did not capture meaningful word combination which could be called units of meaning. It would be odd to expect humans to produce meaningless responses like *of* → *the*. Second, 2-word collocates do not always form units of meaning: very often they are only parts of extended units of meaning to identify which we need to take much more context into account. The procedures Mollin carried out do not allow for more abstract associations either: if a 2-word collocate does not match a stimulus-response pair verbatim, it is possible that they match at the level of colligation of semantic preference.

Also, Mollin admits that indeed as we know language patterning depends on register, but points out that WAT has not been discussed in terms of the register it may belong to, and therefore she does not see this as an obstacle for implementing her research design. Yet, in principle the context of situation may matter for WA responses elicited just like the specificity of discourse determines language use. And just how the context of situation and the selection of stimuli may influence the responses, we have seen from Jung's example.

Another study, (Michelbacher et al. 2011), examines asymmetry in corpus collocations and finds human associations helpful and relevant in this respect: in the word combination *Christmas decorations, Christmas* would predict *decorations* better than *decorations* would predict *Christmas,* both in a corpus and a WAT.

So, we may conclude, the seemingly problematic nature of word associations again boils down to the research questions which can and cannot be asked when WA stimulus-response pairs collected in some certain way are used as data.

'Priming' effects

One common observation made by WA researchers, partly explaining persistent interest in WAT, is that despite considerable variation each stimulus word elicits a large number of identical responses. As Fitzpatrick insightfully notes:

> [the] response word is, therefore, the product of a tension between two influencing entities: the cue word and the respondent. It follows that if the cue word were the only influence, all responses might be identical; if the respondent were the only influence, all responses would quite possibly be different. In reality, of course, some cue words have a stronger 'influence' on the response word than others. (Fitzpatrick 2007: 322-323)

In this respect, one word association pattern which does not get enough attention, to my mind, is the fact that, as retrospection interviews show, very often the response is directly related to the respondent's prior language experience, that is, we see a priming effect.[49] For example, in the study reported by Fitzpatrick (2006), *manual* triggered *vacuum cleaner* because the respondent had been looking for the vacuum cleaner manual that morning, or *liberal* prompted *inaccurate* because the respondent thought a liberal interpretation of the bible was inaccurate. In the present study *pursuit* was associated with *trivial pursuit* because that was the name of the game the respondent had played recently, or *target* prompted *language* because the respondent worked at a translation agency and heard the combination very often. Along the same lines Albrechtsen et al. (2008) mention that "most of the associations that initially seem odd or extremely idiosyncratic turn out to be lexical combinations that are used widely as creative collocations; for instance, in the form of names of a musical group, titles of songs, lines in poems or simply as brand names for different types of products" (46). In other words, shared linguistic experiences might bring in repeated associations, while more individual features of language exposure, like the recency of a particular priming, might result in responses which might look more idiosyncratic at first glance.

Conclusions from previous WA research for the present study

So far we have looked at what different researchers wanted to find in word association responses and how they applied the method. A number of important conclusions can be drawn from this review for the present study.

The design of the WAT, the setting in which it is administered and the selection of stimuli – all can influence the word associations collected. Any specificity of a respondent's profile of word associations, e.g. seeming preference for syntagmatic responses, may be a result of following a strategy in providing WA responses. Direct comparisons of WA responses to some norming or canonical responses may not give any insight into the sanity or

[49] See Section 2.8 for the definition with which 'priming' is used in this study.

language proficiency of the respondent. Word associations alone, without anchoring to a different type of data from the same respondent may be too open to subjective interpretation. At the very least, WA responses need to be complemented with the respondent's introspective comments; otherwise the risk of a mistaken interpretation and categorisation of the responses is very high.

In this study, word association responses are anchored to both the respondents' recent, attested language usage and recent, attested language experience, strengthened by their interview responses. To ensure the comparability of WA responses with the usage data and the priming data, WAs were collected in the same discourse settings, as it were. While WA responses have never been considered register-specific in the same way as language patterning is, as we have seen from Jung's example, the context in which a WAT is administered may have a direct effect on the respondent's interpretation of the task and the responses which are elicited. In the present study, the respondents were steered towards academic discourse in several ways. First, the WATs were administered during the meetings where the students' theses drafts were discussed. Second, all the stimulus words were taken from the students' own writing (see the next section for a detailed discussion of how the stimuli were selected).

The students were not instructed to give associations from their theses or academic field of interest, they were not even told that the words they see in their WATs are taken from their own writing, but all of them noticed this fact at some point. Here is a comment from one of the students, when I asked her whether she thought the fact that the stimuli were taken from her texts influenced her associations in any way:

> … probably in some cases like if I would see *finite* in some word list that is not necessarily, that I would not have realized that all the words come from my text, then I could have written something else there maybe. But when I start reading these words, then I sort of get into the feeling of my text, then of course that is the first thing that comes into my mind. (Kaisa)

This way a WAT still remains a decontextualised task since it is stripped of the requirement to communicate a meaning, but at the same time it is embedded into a discourse, just as all natural language use is, and therefore might get closer to the common psycholinguistic processes involved in natural language production.

The present study: selection of stimuli

The selection of stimulus words required a methodology of its own to be developed.[50] In line with the previous research, the initial experimenting with the WAT in the beginning of the study showed that WA responses divide into two groups:[51] meaning-based (or paradigmatic)[52] and syntagmatic. Meaning-based responses seem to involve an interpretation of the meaning of a stimulus word and often reflect its semantic relations: these responses can for example be synonyms, antonyms, or meronyms (see Section 6.1.1 for a detailed description of responses which were categorised as meaning-based). In giving syntagmatic responses, on the other hand, the respondents do not appear to exploit the semantic relations of a stimulus word but instead supply words which can go together with it, precede or follow it in text (see Section 6.1.2). However, it was also noticed that not only some stimulus words seem to "have a stronger 'influence' on the response word than others", according to Fitzpatrick's (2007: 322-323) observation, but there may be a more complex interrelationship between the type of a response elicited and the properties of the stimulus word. Therefore in the selection of stimuli, it was deemed important to make sure that different 'types' of words are represented in the lists of stimuli. Later this measure allowed me to correlate the behaviour of words in usage and in word association tasks.

The first question to solve was to decide what types of words need to make their way to the word association tasks to achieve a good representativeness of different types of words. First of all, we know that high-frequent and low-frequent words behave differently. Highest-frequent words are very high-frequent (Zipf 1935) and for this reason may be expected to differ in their behaviour from all kinds of other words. They include functional words (e.g. *the, and, of*), but also lexical words which participate in a large number of different patterns, like the verbs *make* and *take*: in Sinclair's terms they are for that reason often delexicalised (see Section 2.4); in dictionary terms, they are polysemous, in that they have many different senses. To make sure that both high-frequent (including highest-frequent) and low-frequent

[50] In the research literature, notably Tess Fitzpatrick (2006, 2007) argues for the importance of compiling a list of stimulus words in a principled way. She herself uses Coxhead's AWL to select the stimuli from. One reason for this is that AWL excludes 2000 most frequent words in English and therefore suits Fitzpatrick who prefers to avoid high-frequency words since as some studies show they tend to elicit more predictable responses.

[51] Some of the responses were also form-based or clang responses, i.e. the responses which are not related either to the meaning of the stimulus word or its collocational behaviour, but the number of them was insufficient and therefore not looked into in this study.

[52] Classification of word association responses into paradigmatic and syntagmatic is conventional in word association research. An association is considered paradigmatic when it belongs to the same grammatical class as the stimulus word. However, in the classification scheme I have adopted there is no requirement for this condition to be fulfilled, therefore this type of associations is called meaning-based (after Fitzpatrick 2006), even though most the associations in the category are comprised of e.g. synonyms, antonyms, meronyms representing standard paradigmatic relations.

words are selected as stimuli, two methods were used. First, I retrieved wordlists using AntConc. Such lists specify the frequency of occurrence for each word used in a text, which can be the whole production corpus of a student or just one draft, and its rank, and thus, allows one to choose words of different frequency of occurrence for a particular text. Second, I generated Lexical Frequency Profiles (Laufer and Nation 1995), again, for either the whole corpus or some part of it. The programme divides the words used in the text into four bands: (1) those which belong to the first 1000 most frequent words in English, (2) those from the second 1000 most frequent words in English, (3) academic vocabulary (the programme compares the text against Coxhead's AWL) and (4) those not in the lists, i.e. those words which are less frequent that the 2000 most frequent words in English and do not belong to the 'academic core', that is words which are relatively low-frequent. Choosing words for WATs from these four different bands allowed me to ensure certain variety.

But the continuum between high-frequency and low-frequency is just one of the dimensions on which words differ. Some words can be part of extended patterning, from a simple collocation to a more complicated pattern consisting of variable verbatim and abstract components and some words make meaning relatively independently of other words. To ensure that words with syntagmatic associations were included in the pool of stimulus words, I also generated automatic lists of 2- to 5-word n-grams with AntConc. Words from the lists were then chosen fairly randomly, but using common sense: it was important that words of a certain type are incorporated into WATs, it was not that important which words in particular were taken. To give an idea of how the selection process worked in practice, Table 4.4 shows an extract from Kaisa's list of 2- to 5-word n-grams sorted by their frequency with words selected for her WATs in bold:

Table 4.4 An extract from Kaisa's list of 2- to 5-word n-grams sorted by their frequency

Rank	Frequency of occurrence	2- to 5-word n-grams
1	53	of the
2	36	in the
3	23	historical **linguistics**
4	20	to the
5	16	and the
6	15	can be
7	12	the cognates
8	12	the proto
9	11	of a
10	10	on the
11	10	the data
12	10	two level

13	9	**based** on
14	9	cognate **recognition**
15	9	from the
16	9	have to
17	9	it is
18	9	level rules
19	9	**sound** changes

There is a third perspective to be taken into account too. Scholars working with academic discourse and teachers of English for academic purposes also distinguish academic vocabulary, which is evident from the interest in developing academic word lists such as AWL (Academic Word List, Coxhead 2000) and AKL (Academic Keyword List, Paquot 2010) and academic phrase lists such as AFL (Academic Formulas List, Simpson-Vlach and Ellis 2010). This type of vocabulary is especially important for this study, given the context of data collection and the type of usage data collected. In theory, it is academic vocabulary which was likely to be in the process of development for these students as they were writing their Master's theses. Therefore, as has already been mentioned, it was considered necessary to take some words from the academic vocabulary band of the Lexical Frequency Profile (LFP). Yet this measure was not considered sufficient to capture the students' field-specific lexis since AWL was compiled so as to represent academic vocabulary shared between different fields of science. For this reason, "Not in the lists" band from the LFP and keyword lists generated with AntConc were also examined for field-specific terminology.

To sum up, lists of stimuli for the WATs were comprised of words taken from different frequency bands of a wordlist or an LFP, components of 2- to 5-word n-grams, academic vocabulary from an LFP's AWL band and keywords. The words from these different lists were then carefully shuffled, so that words of the same kind did not directly follow each other since it could trigger chaining or (short-term) priming effects. For example, if words with clear syntagmatic associations followed each other, a test-taker could adopt a strategy of generating collocations for all the rest of the stimuli down the list. I also kept the order in which the stimuli were presented to the test-taker and responded to and numbered them. This way I was able to check the possible priming effects, when the same response was given to more than one stimulus. The words were presented to the test-takers in their original grammatical form, i.e. in the form they appeared in the text, to encourage the respondents to treat the stimulus words as purely linguistic items rather than reach out for the conceptual (or psychological, as in Jung's experiments) sphere lying behind them.

Administration

A list of stimulus words for a WAT was prepared every time a student submitted a new draft for her thesis and a meeting was scheduled. At the meeting, the WAT was administered first, before we started any discussions since everything that was said could prime the students in their responses. The task was formulated in the following way (see Appendix A for an example of a typical test): "Please write down the first word(s) you think of when you read each of the words listed". Stimulus words were listed in a column, and the students had to write their association in the second column next to each stimulus word.

As such, the students read the stimulus words rather than heard them and had to write their responses rather than speak them. This is important since as we have seen from Jung's examples, the context in which a WAT is administered can have a profound influence on the test-taker's responses. In this study, WAT had to approximate the process of writing a Master's thesis as a far as possible, in terms of the stimuli chosen and their ordering, the context and mode of activity. Also, we know from corpus linguistic studies that language is register-specific (see e.g. Biber 1988; Biber et al. 1999) which means that lexical patterns we tend to see written and lexical patterns we tend to hear spoken are also going to be different. Psycholinguistic processing when speaking a word and writing it might also be significantly different. It would be an extremely interesting question to investigate whether written and spoken WATs can lead to different kinds of outcome, but for this study it was safer to follow the principle of approximating the data collected for C1 in the design of the test to make the comparison legitimate and eliminate any possible variables.

The response times were not measured, since these were simple paper-and-pencil tests. However, the approximate time which went on completion of a whole test was recorded to get a more general picture of the degree of spontaneity with which the responses were given. Table 4.5 lists the WATs taken by the students with the date of administration, number of stimuli, number of responses, the approximate time which was spent on responding to the task and the calculated average time spent on responding to each of the stimuli. The WATs marked in bold were selected for the analysis described in Chapter 6.

Table 4.5 WATs the students have responded to

Student	N of WAT	Date	N of stimuli	N of responses	Timing (approx.)	Timing per word
Kaisa	1	250609	65	65	no time	
	2	140909	103	103	12 min 55 sec	7.5 sec
	3	300909	101	101	12 min 25 sec	7.4 sec
	4	021209	109	109	13 min 50 sec	7.6 sec

	5	**130410**	**135**	**135**	**13 min 00 sec**	**5. 8 sec**
	6	**300410**	**121**	**121**	**9 min 20 sec**	**4. 6 sec**
Hertta	1	220910	103	103	7 min 00 sec	4.1 sec
	2	011010	92	92	9 min 15 sec	6 sec
	3	**141010**	**120**	**120**	**11 min 30 sec**	**5. 75 sec**
	4	151210	115	115	8 min 40 sec	4.5sec
	5	050111	119	119	10 min 50 sec	5.5 sec
Maisa	**1**	**140110**	**110**	**91**	**12 min**	**6.5 sec (7.9)**[53]
	2	031110	138	122	11 min 20 sec	5 sec (5.6)
	3	281210	111	89	7 min 40 sec	4.1 sec (5.2)
	4	280211	100	68	6 min 25 sec	3.85 sec (5.4)
Linda	1	150410	116	116	16 min 55 sec	8. 8 sec
	2	030510	104	102	12 min 30 sec	7.2 sec
	3	050510	118	118	12 min 20 sec	6.3 sec
	4	**100510**	**116**	**116**	**11 min 00 sec**	**5.7 sec**
Nora	**1**	**060510**	**122**	**122**	**10 min 20 sec**	**5 sec**
	2	190111	122	122	9 min 50 sec	4.8 sec
	3	030211	106	106	5 min 40 sec	3.2 sec
Total/Mean			2446	2355		5.5 sec

After the student finished responding to the task, she was asked to read each of the stimulus words and her response to it (used later to confirm my understanding of the handwriting) and briefly comment on whether it was hard to come up with the response, why she thought she gave that response and whether the word was relatively familiar or unfamiliar to her. Answers to the first question were intended to provide some insight into the automaticity of the response: if the response was hard to arrive at, it probably required some conscious thinking. Answers to the second question were often very helpful in interpreting the association: for example an association like *actually* → *love* could not be classified as syntagmatic unless it was documented that it was in fact a movie title. The third question was designed to obtain some information on the stage of acquisition or level of entrenchment, that is whether the item was relatively new and just starting to be productively used or, vice versa, an old item with confident usage patterns.

Limitations

One of the main limitations of the WAT for this study is the fact that it is difficult to argue on the basis of WATs alone that there are some responses which are produced or constructed using explicit knowledge and some responses retrieved from implicit memory by syntagmatic

[53] The time spent per one stimulus rather than one response. The time spent on one response is in parenthesis. divided by the number of stimuli, not the number of responses because she spent time on all of the stimuli even if she could not come up with a response.

association. Determining how automatic a response was could only be done indirectly, through the analysis of the respondents' retrospective comments, since the responses were not timed. Reaction time information could have helped, at least to divide the responses in two groups: fast and slower, but the design of the study would have had to be much more complicated. It would have to be a laboratory experiment rather than a simple paper-and-pencil task: the stimulus words would have to be presented on the computer screen and the respondents would have to say their associations aloud to be recorded by the computer to avoid the interference of the typing speed. Then the computer programme should have been able to generate a list of stimulus-response pairs automatically because it was important to discuss the results of the task with the respondent immediately after the task.

There are also problems of possible (short-term) priming effects and adoption of strategies in taking the task, even though they were alleviated by carefully arranging the stimulus words and keeping the ordering of responses to check for possible identical responses to different stimuli. Also, the response a respondent writes down is not necessarily the first response that has come to his/her mind. The respondents may not have the means to express that first response at all, for example in case it is merely a vague semantic association.

Due to all these factors, it is necessary to thoroughly analyse a large number of word association responses for each individual respondent to discover any possible trends. In this study 350 to 634 word association responses were collected from each of the five participants (2,355 in total) in 3 to 6 tests administered. WA responses from one WAT for each participant were taken for detailed qualitative analysis described in Chapter 6. The rest of the WATs served as evidence of the fact that the respondents did not produce strikingly different associations each time they took the test making it reasonable to take one WAT from each respondent as a sample of her WA responses. To test this assumption quantitatively, two WATs were analysed for one of the respondents, namely Kaisa, showing that the observed tendencies in one WAT are similar to the observed tendencies in the second. Also, all word associations were used as a database to search in when comparative data was needed for certain individual cases. With several tests for each respondent collected, it was also possible to examine whether the responses to the same stimuli change over time or not.

4.2. Methods of analysis

Both usage data and word association data are very rich, and therefore it is very easy for a researcher to fall into the trap of finding only those examples which support the hypotheses. At the same time it is not possible to analyse each word in each corpus and database from three different perspectives: its usage patterns, its source of acquisition and its representations in the mind. Thus, to make the analysis comprehensive and avoid the danger of cherry-picked examples, several steps were taken. First, units of meaning were analysed according to several explicit criteria (see Section 4.2.1). Second, informed decisions were taken as to how a concept of a unit of meaning can be operationalised to facilitate automatic retrieval of at least a representative proportion of all units of meaning occurring in usage data (see Section 4.2.2). Third, the comparisons between usage data and word association data and between usage data and priming data were completed independently. These two studies are described in Chapter 5 and Chapter 6 respectively. Examples from the three different types of data are finally put against each other in Section 6.3.8. The analysis in each of the studies was broken down into simple stages and procedures repeated for each data set comprised of one student's C1, C2 and WAs (see Section 4.2.3). Five data sets were analysed in total to see whether the observations in each data set support each other. Fourth, qualitative and quantitative analyses were combined in each study (Section 4.2.4). In the subsections that follow, all of these steps will be elucidated in detail.

4.2.1. Analysing units of meaning

Since the overall aim of the study is to examine L2 users' operation on the idiom principle, in both analysis chapters, Chapter 5 and 6, I apply the model of a unit of meaning to the analysis of the observed patterning. The theory behind the model is discussed in Chapter 2.

In principle, to establish that a pattern constitutes a unit of meaning, one must show that it is produced on the idiom principle because it is when a pattern becomes produced on the idiom principle that it undergoes a meaning-shift. This is problematic because the argument threatens to become circular on the one hand, and on the other hand there are no direct means of showing that a stretch of language is produced by co-selection or by syntagmatic association from implicit memory. Even though direct evidence cannot be obtained, there are forms of indirect evidence that can be used instead.

In order to show that a pattern constitutes a unit of meaning, I applied the following criteria: (1) recurrence, since if a pattern recurs, it is unlikely to have been constructed from scratch every time, (2) compliance with the model of a unit of meaning, since the components which comprise units of meaning have already been established in previous research, (3) the integrity of its communicative function, i.e. its semantic prosody, and the consistency with which it is used, (4) an associative link between its components as shown by a WAT since this is taken as an evidence of its being represented in the mind in a holistic form. Ideally all of these criteria should be satisfied, but it seems if the data provides evidence for only three of them, it is still possible to postulate a unit of meaning for the following corresponding reasons: (1) a pattern can be produced on the idiom principle without being repeated even twice since C1 does not contain everything a language user has ever heard, read, said or written, (2) it is possible that not all the components of a unit of meaning have been identified in previous research, (3) semantic prosody is an obligatory component of a unit of meaning, but its interpretation cannot be automatised and is therefore fully dependent on the interpretation of the researcher, (4) a syntagmatic association may not happen to be retrieved in a WAT even if it exists (the reasons for it are discussed in Chapter 6).

The data for the analysis of the first three criteria was retrieved by simply generating concordances for a certain query word from a corpus of usage. For this purpose either AntConc (Anthony 2007) or ConcGram (Greaves 2009) were used. Yet, the more difficult task is to show the extent to which one is operating on the idiom principle. Since it is not possible to analyse every word produced in the exhaustive fashion outlined, it is necessary to think of possible ways of operationalising the concept of a unit meaning. The next section elucidates the solutions developed in this regard.

4.2.2. Operationalising units of meaning

As explained in Chapter 2, a unit of meaning can be at minimum comprised of just one word. This is a limiting case but important for the argument because it removes the boundary between single-word and multi-word lexical items and puts meaning rather than orthographic conventions at the heart of the matter (see the discussion in Section 2.3). Yet, we are interested in extended units of meaning, or units of meaning comprised of more than one word.

Unfortunately it is not possible to retrieve a list of all extended units of meaning occurring in a corpus automatically for three reasons: (1) in identifying co-occurrences, corpus software has to rely on frequency as a criterion, so the list retrieved is always a list of recurring co-occurrences, and the ones which occur only once remain unnoticed, (2) corpus software cannot judge which of the co-occurrences are meaningful, and the solutions which are used to filter the lists give only approximate results, (3) corpus software can detect only verbatim co-occurrences, but not abstract associations such as semantic preference and colligation.

One of the tools which comes close to the task of automatically retrieving a list of all possible patterns appearing in a corpus is the phraseological engine ConcGram (Greaves 2009). Its advantage is that it retrieves co-occurring combinations of words irrespective of their positional or constituency variation, i.e. a co-occurring word can occur to the right or to the left from the origin word (i.e. the search item, this terminology was mentioned in Section 2.2), right next to it or including one or more words in between. Therefore, for example, it is able to detect a co-occurrence of *undergone* with *changes*, as Example (4.1) illustrates.

(4.1)

1 from each other a long time ago, they have **undergone** many **changes** and the cognates might look very
2 millennia ago. Finnish, on the contrary, has **undergone** several **changes** during that time, many of which
3 by separate sound **changes** that the words have **undergone** independent from each other. The

(Kaisa, C1, concordances for the concgram UNDERGONE/CHANGES)

The span can also be customised, so since in this research study a conventional internal span of 4 was used, i.e. the program was asked to show all the words which occur more than once to the right or to the left from the origin with up to four intervening words in between, the occurrence in the third concordance line was also retrieved as there are exactly 4 words between *changes* and *undergone.*

However, ConcGram does not have an option of lemmatising the search in creating a concgram list automatically, so the related concgram UNDERGONE/CHANGE will be listed separately, as the concordance lines in (4.2) show.

(4.2)

1 that it is more likely for one language to have **undergone** a **change** than for many languages to have
2 [1] (*p), because only Hungarian would have **undergone** the **change** *p > f while Finnish and Udmurt would
3 have retained p and only Hungarian would have **undergone** the **change** p > f. The next guideline to help

(Kaisa, C1, concordances for the concgram UNDERGONE/CHANGE)

The instances of the same collocation displayed in (4.3) will not be captured at all.

(4.3)

1 . sometimes so strong that it keeps the word from **undergo**ing otherwise regular sound **change**s.
2 too, initially had only one sound, k, which **underwent** the **change** k > h before front vowels. Table 1:

3…contact, both languages, Sumerian and Akkadian, **underwent** a lot of **change**s, visible for example in

(Kaisa, C1)

Yet, it is possible to lemmatise the search when searching for concgrams of specific words, either by keying them in as an Inclusion List or by using the Concordance Search Dialog. It is also possible to search for all the instances of a specific concgram by selecting the appropriate parameters in the ConcGram Search Dialog: for example treat each of the two words in a concgram as a word or prefix. It is necessary to keep in mind though that this measure will for example retrieve all the instances of *undergo, undergoing, undergone* and *undergoes* but not *underwent.*

Another problem is that an automatic list of all 2-word concgrams retrieved even from a small corpus like C1 is really huge. Table 4.6 shows the number of concgram types generated for each corpus in comparison to the corpus size.

Table 4.6 All 2-word concgrams (types) and significant concgrams (types and tokens) for each C1

Student	C1 size	All 2-word concgram types, C1	Significant concgrams (T≥2 & MI ≥3), types, C1	Significant concgrams, tokens, C1
Hertta	39,449	19,828	464	4,523
Kaisa	21,887	14,019	347	3,283
Linda	37,641	20,085	524	6,004
Maisa	21,501	11,183	260	2,451
Nora	14,843	8,438	108	1,414

The third column shows the number of all co-occurrences of any two words with no more than 4 words in between occurring in the corpus at least twice retrieved from corpora without any cut-off points or exclusion lists. It is not difficult to calculate that the number of total instances of these concgrams would cover the number of running words in a corresponding corpus more than twice. The reason for this is that one and the same running word can participate in more than one 2-word concgram, that is, it can be in fact a 5-gram but in a list of 2-word concgrams it would be represented by 10 2-word concgrams. It is not feasible of course to work with these numbers, and some kind of filtering is indispensable especially since many meaningless word combinations are retrieved this way.

The two statistical values standardly used in Corpus Linguistics to determine significant co-occurrences, and the ones actually available in ConcGram, are t-score ≥2 and MI value ≥3 (McEnery et al. 2006: 56-57; Hunston 2002: 71-72). These values can be used as cut-off points and applied to the list of concgrams iteratively. To avoid manual filtering as far as possible, which is very labour-intensive and time-consuming, it was decided to apply both statistical filters. The concgrams on the outcome lists are then called *significant concgrams*. The fourth column in Table 4.6 shows the substantially reduced numbers of such significant concgrams.

There certainly are some limitations to this approach, which are best visible through examples. The concgram GRAVE/RESERVATIONS occurring 2 times in Kaisa's data set has the t-score of only = 1.4, although its MI value = 10.1, so it is filtered out from the final list. On the other hand, none of the concgrams representing the pattern *it is possible to* occurring 19 times (IS/POSSIBLE, IT/POSSIBLE, POSSIBLE/TO) reaches the MI value threshold = 3. EXAMPLE/OF (19 occurrences) is not significant by either of the measures (t-score =1.8; MI = 0.67). So even though all of these three concgrams form meaningful patterns, they are not represented on Kaisa's list of concgrams. However, the purpose of the study is to find out to which extent L2 users operate on the idiom principle or, in other words, which proportion of their text is produced on the idiom principle rather than to calculate how many extended units they produce exactly. Therefore, the procedure of generating an automatic list of concgrams filtered with statistical tests seems permissible. Yet, it is important to bear in mind that the number of significant concgrams submitted to analysis in this study is not the number of extended units used in C1, neither can the share of text in running words they form be compared with the earlier counts of the pervasiveness of the idiom principle in language production (Erman and Warren 2000 report about 50%, Altenberg 1998 - even 80%, while Dąbrowska 2004 points out that even this can be an underestimation). This number of significant concgrams can only be considered as indicative of the general trend.

The fact that sometimes the analysis relies on automatic extraction and sometimes on careful qualitative examination also explains the alternating use of phraseological terminology. In most of the cases when working with automatically generated concgrams (or n-grams, see Section 4.2.3), it is obvious that the patterns in question are multi-word units. However, in quantitative analyses and other cases where I do not have a possibility to perform a deeper analysis according to the criteria of a unit of meaning, I refrain from calling these patterns extended units of meaning and use the term a multi-word unit instead. But at the stage of qualitative analysis examples of MWUs taken from different categories are inspected in terms of a model of a unit of meaning, and it is assumed that these examples are representative of other cases in their category. That is, categorisation is used as a sampling method.

4.2.3. An overview of the procedures

While both chapters, Chapter 5 and Chapter 6, explore L2 users' operation on the idiom principle, the more specific division of labour between them is the following.

The aim of Chapter 5 is to answer the question whether the idiom principle is available to L2 users and if it is, to what extent. Therefore, in this chapter I bring together different kinds of evidence which are able to illuminate L2 users' operation on the idiom principle, or lack thereof. In particular, (1) I generate concordance lines for several query words and apply the model of a unit of meaning to their usage patterns as described in Section 4.2.1. Then (2) I explore whether such kind of patterns were constructed or learned from exposure on the idiom principle by comparing students' production corpora (C1) to their priming corpora (C2). Technically, this is done by retrieving an automatic list of concgrams from C1, filtering them using statistical thresholds MI = 3 and t-score = 2, as described in Section 4.2.2, and automatically comparing them to the patterning found in C2, which is an additional functionality ConcGram offers. The automatic comparison of C1 concgrams to C2 shows the proportion of these C1 concgrams which also occur in C2, that is, answers the question how many of the patterns students use also occur in experts' writing.

Since the resulting figure cannot in itself reveal whether the overlap found should be regarded as large or small, I take two further steps. First, in addition to the comparisons within one data set, I carry out comparisons across data sets. That is, I compare one student's C1 concgrams to a different student's C2 to see whether the extent of overlap will stay the same. Second, I analyse the concgrams qualitatively. For this purpose, I divide C1 concgrams on the list into those overlapping with C2 i.e. *Matching* and those which are unique to C1, i.e. *Non-matching*. The analysis of the examples from the Matching concgrams can tell us how closely the students reproduce the patterns they encounter in the expert writing of their field and what kind of phenomena are observable when the pattern is overlapping. The concgrams in the Non-matching category can shed light on the ways in which students' usage patterns diverge from the ones they were exposed to.

As can be surmised, the decisions about operationalisation of units of meaning discussed in Section 4.2.2 are crucial here because the patterns that are retrieved and analysed wholly depend on them. We have seen that there are two major problems caused by operationalising units of meaning as significant concgrams: not all the units of meaning are retrieved and not all the units of meaning which are retrieved are meaningful. The first problem cannot be resolved. But the second is addressed at the stage of qualitative analysis when all the Matching and Non-matching concgrams are supplemented by concordance lines.

Again since C1 corpora for all the five students generate 1703 significant concgrams, it was not feasible to analyse all of them qualitatively. For this reason I decided to focus on two data sets, that of Kaisa and Maisa, comprising 607 significant concgrams in total.

To complement the analysis undertaken in Chapter 5, in Chapter 6 I explore how the idiom principle works: postulating that the idiom principle is a psycholinguistic mechanism, I investigate whether the model of a unit of meaning has psycholinguistic reality. To do that, I compare each student's usage data against her word association responses. To be more specific, for each stimulus word, the response it elicits is compared to the pattern it participates in in usage.

Just as in Chapter 5, here too the list of significant concgrams was used to represent usage patterns. However, in contrast to Chapter 5, this fact is not crucial for the outcome of the analysis. The list of significant concgrams was used only as an aid to speed up the process of comparing WA stimulus-response pairs to usage patterns. When a list of WA stimulus-response pairs and a list of concgrams are conflated in one Excel document and ordered alphabetically, the possible matching concgrams are located right beneath the WAs, so there is no need to query each stimulus word in C1 and see whether the emerging usage pattern matches the response word. Since we know that not all meaningful units end up on the list because of the not always effective statistical thresholds, another list of co-occurrences was generated: 2- to 5-word n-grams produced with AntConc (Anthony 2007). Their advantage is that the requirement of contiguity is itself a very good filter, and thus, they do not need to be further filtered with any statistical tests. However, as was already mentioned, both lists were only used as an aid: in all unclear cases and all cases where there was no concgram or n-gram retrieved, the WA stimulus word was used to generate concordance lines in order to inform the analysis.

To get an overall picture of the relationship between WAs and usage patterns, the ways they compare to each other were categorised. A word association stimulus word can elicit a syntagmatic (S) or a meaning-based response (M). In usage, this word can either participate in a certain pattern (MWU), an extended unit of meaning, or function independently (No MWU).[54] When a stimulus-response pair is compared to the usage pattern of the stimulus word, it can either match it or not (Matching/Non-Matching). Different scenarios result in five groups:

(1) Matching MWU S-responses,

[54] See Sinclair 1991: 71 for a similar division of usage patterns into *independent* and *dependent*.

(2) Non-matching MWU S-responses,

(3) Non-matching MWU M-responses,

(4) No MWU S-responses,

(5) No MWU M-responses.[55]

Division into the categories allowed me to calculate whether the interrelationship between WA responses and usage was statistically significant. Also, it shows how many of the responses matched the usage out of those which could have been matching, since, as the No MWU category indicates, it is possible that there was no pattern to match in the first place.[56] Third, by taking examples from each category, it was possible to achieve a comprehensive analysis of all the data. And fourth, the analysis of examples from different categories is able to shed light on different aspects. For example, by analysing Matching MWU S-responses, it is possible to see what kind of syntagmatic responses there are: in particular, whether these associations can only be verbatim or they can be abstract too. Non-matching MWU M-responses can give an insight into the reasons behind the cases where an expected syntagmatic response is not given.

This section has given an overview of the procedures undertaken in each of the chapters. Next section will summarise the unifying principles on which these procedures were structured. The details of the procedures which are tightly interlinked with the analysis itself will be discussed in the corresponding Chapters, 5 and 6.

4.2.4. Combining qualitative and quantitative analyses

The two analyses, reported in Chapter 5 and Chapter 6 respectively, are structured symmetrically. In Chapter 5, production corpora are compared to language exposure corpora. In Chapter 6, production corpora are compared to word association stimulus-response pairs. So each production corpus constitutes a pivotal point around which the analysis is built in each case.

In both chapters, the deeper qualitative analysis is built on initial quantitative investigation. In the first case, the quantitative result is able to tell us to what extent C1 and C2 corpora overlap in terms of multi-word unit patterning, in the second case whether there is a relationship between usage patterns and word association responses. The resulting figures

[55] While Matching MWU M responses are in principle thinkable, in reality they do not exist since if a response matches a MWU pattern, it is categorised as an S-response because of the definitions of S- and M-responses used in this study.

[56] Since WA stimuli were selected randomly (or so that they would test the behaviour of different types of words), they may represent different types of words in different proportions. Therefore, e.g. the proportions of S-responses, M-responses or even Matching MWU S-responses cannot tell anything on their own. Only their interrelationships can be indicative of certain possible tendencies.

are able to show whether further qualitative analysis is justified or whether any further questions are irrelevant because there is no relationship between the two types of data.

The qualitative analyses are then based on the quantitative outcomes. In the first case there is a number of C1 co-occurrences which are found to be matching to C2 and a number of them which are not. By further categorising the Matching and Non-matching concgrams or co-occurrences into smaller groups by some kind of critical feature which unites them and analysing examples from each of the resulting categories it was possible to cover all the data exhaustively. The categories in the outcome are then entirely data-driven. In the second case, classification builds on the binary oppositions detected in the usage data, WA responses and the way they compare to each other, which resulted in 5 categories as described above in Section 4.2.3.

4.2.5. Using the BNC

In examining students' usage patterns in comparison to their priming language, not only their C2 corpora are used but also the BNC as a general-purpose corpus. This is done on the following grounds. As pointed out in Section 4.1.4, C2 corpora were compiled so as to represent the students' field-specific priming language. At the same time, the language represented by general-purpose corpora like the BNC is part of the students' language exposure too. Therefore, in all cases when a C1 pattern is found in C2, the BNC or in both corpora, it can be argued that it is learned from exposure. However, the more subtle field-specific patterning is much less likely to be introduced in any language learning reference materials such as dictionaries and thus is unlikely to have been learned explicitly through formal instruction in classroom contexts, but instead is more likely to have been learned implicitly through exposure. Implicit learning of phraseological patterning through exposure is likened to learning on the idiom principle. Therefore, the examples of learned field-specific patterning are especially valuable for the argument and are given more focus.

The BNC, which is representative of Standard (British) English, was also used when there was a need to position the observed patterns in a wider context. It is used as complementary data in correlating WA patterns to usage patterns in Chapter 6 too.

4.2.6. A note on notation

Italics are used for *word-forms* and word association *stimulus → response* pairs which are also presented with the forward arrow in between, resembling the direction of the association. Small caps are used for LEMMAS and CONCGRAMS. For example in the concgram ASSUMPTIONS/ABOUT, the word *assumptions* co-occurs with the word *about*. In Chapter 5, where concgrams from C1 are compared with concgrams from C2, the frequencies with which a certain concgram occurs in C1 and C2 are given in parenthesis, e.g. ASSUMPTIONS/ABOUT (13/27) means that the concgram occurs 13 times in C1 and 27 times in C2. An asterisk (*) is used for zero or more characters as in the query syntax of the BNC.

The frequencies of occurrences are given in raw numbers rather than normalised. Since C1 and C2 corpora are very small, the numbers of occurrences are so small, that normalising them may only distort the picture. Usually, normalising frequency lets us evaluate the probability of a language user encountering the pattern. Here this probability is already ensured by the fact that the students 'reported' in their reference lists that they were familiar with all the texts comprising C2 corpora. Secondly, since C1 is a production corpus, every occurrence of a pattern is an exemplar, which strengthens the representation of this pattern in the mind. For this reason, it is the absolute frequency rather than the relative frequency which is important. Yet, in case the reader would like to calculate the relative frequencies of occurrence between C1 and C2, and between Kaisa's and Maisa's usage, this is easily done. Kaisa's and Maisa's C1 corpora are almost exactly the same in size: 21,887 and 21,501 words, respectively, and therefore can be compared directly. Kaisa's C2 is exactly 3 times larger than her C1, Maisa's C2 is exactly 4 times larger than her C1, thus, to compare the frequencies of occurrence in C1 and C2, the raw frequency of occurrence in C2 needs to be divided be 3, in Kaisa's case, and 4, in Maisa's case.

4.3. Conclusion

The set task of tracing extended units of meaning in their use, psycholinguistic representation and priming language has required certain unconventional methodological solutions. To observe units of meaning in L2 use, individual corpora of written production were compiled. To find out whether these units used were learned from exposure or constructed from scratch on the open-choice principle by the users themselves, individual priming corpora were compiled. Finally, to investigate whether these units used are also represented in the mind in their holistic form, or in other words, whether the components of these units are connected by syntagmatic association, word associations were collected from the same L2 language users. Figure 4.1 displays these three types of data in overlapping circles which means that every type of data is expected to be interrelated with every other type. The interrelationship between L2 use and priming language is studied in Chapter 5, between L2 use and psycholinguistic representation in Chapter 6. A glimpse of the whole triangle is given in Section 6.3.8.

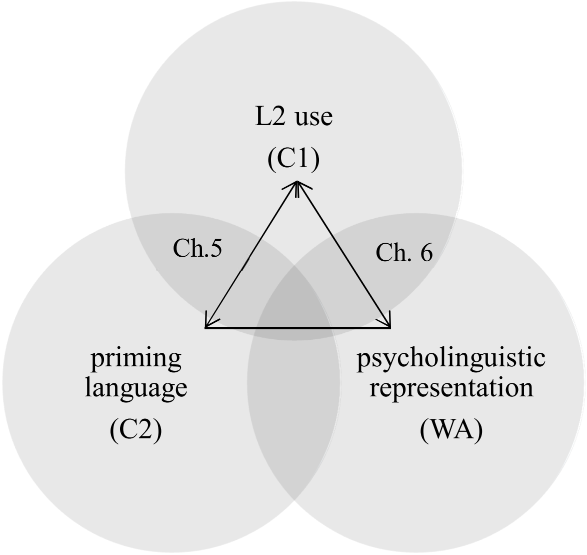


Figure 4.1 Structure of the study

As mentioned at the outset of this chapter, the biggest challenge the study had to face was the richness of data and the necessity to process it in a principled and comprehensive way. As Sinclair notes:

> [L]anguage patterning is too rich for uncontrolled choice; if the researcher can choose only some of the language patterns, then almost anything, and its opposite, can be demonstrated. So if analysis is to be selective then the selection has to be justified and applied uniformly. (Sinclair 2004 [2001]: 116)

Systematicity of analysis had to be ensured at every stage and was introduced by different means: adoption of explicit criteria for the identification of units of meaning, their operationalisation through objective corpus linguistic means, division of analyses into strictly ordered procedures, sampling examples on the basis of data-driven categorisation, combination of qualitative and quantitative analysis. The focus in each analysis chapter constantly shifts to zoom in on certain details and zoom out to embed the refined qualitative analysis in the bigger picture of the typical patterning. Qualitative analysis in its turn gives an insight into the numbers.

Another important feature of the study is its data-drivenness which was pursued at every stage of research starting from the pilot-study and selection of the primary type of data on the basis of experimentation and ending with the data-driven categorisation of data items.

The model of a unit of meaning itself was chosen as the theoretical framework first and foremost because it permits variability and enables data-driven observations as it does not impose strict structure on the data at hand.

While a general account of the data collected and the methods used was given in this chapter, Chapters 5 and 6 provide more details about some of the more elaborate procedures in connection with the analysis itself.

5. Idiom principle in second language acquisition and use: C1 vs. C2

As illustrated in Chapter 3, phraseological competence is usually seen or construed in the literature as a major problem for second language users and learners. However, it is also clear that, to begin with, there does not seem to be a consensus in SLA research on what kind of patterns constitute multi-word units, with the focus tending towards more fixed types of patterning like collocations, lexical bundles or idioms. Second, the conclusions about the unavailability or limited availability of the idiom principle to second language learners are usually made on the basis of direct or indirect comparison with native speaker data which does not necessarily represent the target language competence for these learners or the language they were exposed to, which questions the grounds for the expected similarity between the two kinds of data.

To overcome these problems and obtain a fresh perspective on the phraseological competence of second language users, the following solutions are offered. First, following the argument presented in Chapter 2, I will adopt a wider approach to phraseology, one which accepts adaptive variability inside a unit. I will take Sinclair's model of a unit of meaning as the starting point, but try to keep an open mind for possible other patterns that might emerge from the data. It was suggested in Chapter 3, for example, that in second language use, just like in any other language use, verbatim associations in a unit can be approximated to more abstract associations: that is, a collocation can be approximated lexically/semantically or structurally/grammatically to become a semantic preference or a colligation. I will investigate whether the process of approximation can be observed in the data.

Second, I will try to avoid the methodological limitations of previous studies and examine L2 production in its own right first. I will juxtapose observations of patterning in language use with the language data which most likely served as a source of acquisition for the observed patterns. Both types of language data, L2 production data and source language data, are compiled into individual corpora for each student. While the compilation principles of both types of corpora, referred to as Corpora 1 (C1) and Corpora 2 (C2), are described in detail in Chapter 4, it is important to remind the reader here that the second type corpora are compiled from academic publications so as to represent the language the students were exposed to and were likely to treat as their target when writing their own texts. Representing the kind of language the students would aim at, academic writing is at the same time the kind of language they were less familiar with before starting their Master's level studies. This means that if it can be shown that phraseological patterns from these articles, especially field-specific ones, occur in the students' writing, it is probably freshly acquired language: these would be the patterns which are likely to have been treated holistically from the start and acquired naturally as units.

This chapter is structured in the following way. First, in Section 5.1, I will describe co-selection patterns forming a unit of meaning which are observable in the students' usage data. In Section 5.2, I will compare a list of co-occurrences retrieved from students' usage data to expert writers' texts to see how much overlap there is. In Sections 5.3 and 5.4, the Matching and Non-matching co-occurrences respectively are analysed in more detail. While the analysis of Matching co-occurrences is able to answer the question how closely L2 writers are able to follow expert writers' patterning, the analysis of Non-matching can shed light on the question why some L2 patterns do not match the postulated target language. The two processes which can account for some of the Non-matching patterns are discussed in Section 5.5. The conclusions are drawn in Section 5.6.

5.1. Are the patterns of co-selection observable in the L2 texts?

The data collection described in Chapter 4 resulted in five individual corpora of L2 production. The simplest way to find out whether the authors of the texts - Kaisa, Hertta, Maisa, Nora and Linda - are able to operate on the idiom principle is to search for the usage patterns of certain words in the respective corpora by using concordancing software and see whether they reveal such operation. We can establish that a pattern is produced on the idiom principle if it (1) occurs more than once, (2) conforms to the model of a unit of meaning, and (3) consistently communicates one and the same meaning, its semantic prosody. Let us follow this simple methodology and see what kind of patterns are obtainable this way. We shall start from the more verbatim patterns and then gradually move on to the more abstract ones. For this section, most of the examples will be taken from Kaisa's data set.

In academic writing, fixed co-occurrences of two or more words are very often terminological in nature. A search for the word *training* in Kaisa's corpus of drafts, presented in Example (5.1), yields 6 instances, out of which 5 occur in a phrase *training data* or more specifically *the training data* because the definite article also co-occurs:

(5.1)

1	at misclassified pairs in **the training data** are caused by accidental	kaisa_10_300410.txt
2	and English even without **the training data**. They also suggest that	kaisa_10_300410.txt
3	est consists of two sets: **the training data** and a separate test set	kaisa_summary_210109.txt

4	t misclassified pairs in **the training data** were caused by accidenta	kaisa_summary_210109.txt
5	evaluating the results on **the training** and test pairs, the authors e	kaisa_summary_210109.txt
6	and English even without **the training data**. The method can also be	kaisa_summary_210109.txt

(Kaisa, C1, concordances for the word *training*)

The two words (1) co-occur, (2) form a collocation which is a feasible form of a unit of meaning, (3) are consistently used to express a single meaning. Therefore it is possible to conclude that these three words are likely to be produced on the idiom principle. In this case the argumentation may not look so convincing because we know that terms are fixed units and function as single words, so there is nothing extraordinary in the fact that *training* and *data* co-occur. It is an example from the extreme end of the continuum between more fixed units of meaning and those which exhibit a lot of potential for variability. It must also be noted that in addition to systematic co-occurrence they do not allow any intervening words and are also positionally fixed.

UNDERGO/CHANGE is a different example of a collocation. It is a collocation of two lemmas rather than any specific forms of the verb UNDERGO and the noun CHANGE. The collocation also demonstrates positional and constituency variability (5.2):

(5.2)

1	a long time ago, they have **undergone** many **changes** and the cognates	kaisa_2_140509.txt
2	y for one language to have **undergone** a **change t**han for many language	kaisa_2_140509.txt
3	nly Hungarian would have **undergone** the **change** p > f. The next guide	kaisa_2_140509.txt
4	ly Hungarian would have **undergone** the **change** *p > f while Finnish an	kaisa_2_140509.txt
5	parate sound **changes** that the words have **undergone** independent from	kaisa_3_250609.txt
6	that it keeps the word from **undergoing** otherwise regular sound **changes**	kaisa_3_250609.txt
7	ed via browser and also **undergo version control** in that previous versi	kaisa_5_140909.txt
8	Finnish, on the contrary, has **undergone** several **changes** during that time,	kaisa_8_130410.txt
9	had only one sound, k, which **underwent** the **change** k > h before front vowel	kaisa_2_140509.txt
10	languages, Sumerian and Akkadian, **underwent** a lot of **changes**, visible	kaisa_4_100809.txt

(Kaisa, C1, concordances for the lemma UNDERGO)

Only in line 7 the verb UNDERGO is used with a different object: *version control*. So on the basis of usage it is reasonable to assume that the syntagmatic association between the two words is quite strong.

Moving on to the more abstract patterns, we will now look at a pattern which can be described as colligational. A search for such a general adjective as *important* in Kaisa's data reveals that its usage actually forms a colligation pattern (5.3):

(5.3)

1	se can be disputed, so **it is** first **important to** carefully **present** all the evid	kaisa_5_140909.txt
2	[ref.]. Therefore **it is important to check** the plausibility of the r	kaisa_3_250609.txt
3	e of the comparative method, **it is important to clarify** the essential terms an	kaisa_2_140509.txt
4	ommon ancestor language. **It is important to consider** all evidence available	kaisa_3_250609.txt
5	es by thousands of years, so **it is important to consider** what types of words ca	kaisa_2_140509.txt
6	the availability of cognate lists **important** to computational linguistics.	kaisa_10_300410.txt
7	etween two cognates added. This is **important** because this way it is possible to	kaisa_5_140909.txt
8	lving and open to editing. This is **important** especially in fields like histori	kaisa_5_140909.txt
9	stified. However, even quantity is **important**. If the number of comparable ele	kaisa_1_170309.txt
10	The form contains a field for each **important** piece of information that is relat	kaisa_5_140909.txt

(Kaisa, C1, concordances for the word *important*)

Important in Kaisa's writing is almost always predicative, moreover, more than half of the instances reflect the pattern it v-link ADJ *to*-inf (Francis et al. 1998: 497). In contrast, the synonyms *significant* (5.4) and *essential* (5.5) are almost exclusively used as attributive adjectives:

(5.4)

1	closer inspection, but after a **significant** amount of checking has been do	kaisa_1_170309.txt
2	decision on which of them are **significant** and systematic on their freque	kaisa_9_300410.txt
3	as a regular relation. It makes no **significant** difference on the theoretical	kaisa_10_300410.txt
4	as a regular relation. It makes no **significant** difference on the theoretical	kaisa_4_100809.txt
5	the cognates, it does not make a **significant** difference to the theory to ch	kaisa_9_300410.txt
6	medial sound changes constitute **significant** patterns due to the difference	kaisa_9_300410.txt

(Kaisa, C1, concordances for the word *significant*)

(5.5)

1	od, it is important to clarify the **essential** terms and definitions used in the	kaisa_2_140509.txt
2	work simultaneously. This is one **essential** way in which the two-level model	kaisa_6_300909.txt

(Kaisa, C1, concordances for the word *essential*)

In addition to this, as can be seen from lines 3, (4) and 5, the word *significant* also participates in a pattern MAKE *significant difference.* It can be concluded that all these adjectives are connected for her, but still Kaisa does not use them interchangeably, each word is used in its own pattern.

An example of a semantically abstracted pattern, i.e. a semantic preference, comes from Nora's C1 (5.6):

(5.6)

1	n the discussion forums this might **pose** some **difficulties** since privacy a	nora_2_201010.txt

2	[ref.] calls them - have **posed** new **challenges** and lead to the de	nora_2_201010.txt
3	been done on the research **problem posed** here. In conclusion the literat	nora_1_060510.txt
4	**question** concerning space, the global and local has already been **posed** by	nora_2_201010.txt
5	y subcultures and the **implications posed** by the media are being conducted.	nora_1_060510.txt
6	[ref.]. A spatial concept **poses** the **issue** of boundaries to evaluat	nora_1_060510.txt
7	can be observed. A spatial concept **poses** the **issue** of boundaries to evalua	nora_2_201010.txt

(Nora, C1, concordances for the verb POSE)

As can be noticed from the concordance lines, the verb *pose* co-occurs with a semantically specific group of objects: *implications, issues, problems, challenges, questions* and *difficulties* are posed in Nora's texts. It seems these objects are united by a common semantic feature which can be described as 'matters of concern'. In other words, we can postulate that the verb *pose* has a semantic preference for 'certain difficulties' in Nora's writing.

These examples show that the patterning of a unit of meaning is not atypical for L2 users. The three criteria set for establishing that a pattern was produced on the idiom principle were satisfied in all the cases: (1) the pattern was recurrent, (2) it was structured as a unit of meaning, (3) it was consistently used for one and the same purpose. However, we are confronted with two questions. First, what is the source of acquisition? It is, after all, possible that these patterns are constructed in the first place rather than acquired holistically from exposure. And, second, how typical are these patterns for second language production or to what extent are second language users operating on the idiom principle? I will make an attempt to answer the first question in what follows. As for the second question, it can only be dealt with indirectly since it is not possible to extract an exhaustive list of all the patterns occurring in a corpus (see Sections 4.2.2 and 5.2.1). Thus, it will be answered partly through the first question and partly through the analysis presented in Chapter 6.

5.2. Where do the patterns come from?

Section 5.1 has shown that the patterning observable in L2 texts strongly suggests operation on the idiom principle: it is unlikely that one writer would construct the same pattern from scratch over and over again especially if this pattern fulfils the criteria of co-selection for a unit of meaning. In order to find out whether there is a likelihood that the L2 writers acquired the patterns they are using from exposure, I retrieved a list of co-occurrences from each C1 corpus and compared it to the corresponding C2 corpus. If it can be shown that the patterns of co-occurrences are matching, it is likely that they were acquired holistically through exposure. As discussed in Section 4.2.2, the procedure has certain limitations. On the one hand, it is not possible to retrieve an exhaustive list of all and only meaningful patterns, which makes a complementing manual analysis indispensable. And on the other hand, the retrieved list of all co-occurrences has to be further filtered with statistical measures to make the manual analysis possible. Therefore, before carrying out the comparison, it is useful to evaluate the scope of patterns.

5.2.1. The scope of C1 patterns under investigation

One way to get a general idea of the scope of the patterns on the list of automatically retrieved concgrams from each C1 is to see what proportion of text they cover. Table 5.1 shows each student's corpus size, the number of all 2-word concgrams (types), the number of significant 2-word concgrams (types) together with the total number of their instances (tokens) and the percentage of text these tokens cover.

Table 5.1 (Significant) concgrams in students' production corpora and the proportion of text they cover

Student	C1 size	All concgrams, types, C1	Significant concgrams (T≥2 & MI ≥3), types, C1	Significant concgrams, tokens, C1	% of words participating in significant concgrams, C1[57]
Hertta	39,449	19,828	464	4,523	23%
Kaisa	21,887	14,019	347	3,283	30%
Linda	37,641	20,085	524	6,004	32%
Maisa	21,501	11,183	260	2,451	23%
Nora	14,843	8,438	108	1,414	19%

As can be seen from the table, in C1 the proportion of text covered by significant concgrams is between 19% and 30%. To get a better idea of whether this proportion is large or small, I compared it to the proportion of text covered by significant concgrams in C2 (see Table 5.2). This was considered a worthwhile procedure, even though the percentages of text coverage by significant concgrams between the two types of corpora are not directly comparable because the number of concgrams retrieved from a corpus according to the criteria MI ≥3, t≥2 might depend on many factors, such as: corpus size, the number of authors the texts were collected from, homogeneity of the corpus, specific features of the language of a discipline

[57] To calculate the proportion of the corpus covered by significant concgrams, the total number of significant concgrams is multiplied by 2 to get the number of running words participating in these significant concgrams. The resulting number is rough because, as it was pointed out in Section 4.2.2, one and the same word can participate in more than one 2-word concgram, therefore some of the words can be counted twice or even three times, however, judging by the manual analysis of these lists, it does not happen very often because not all the configurations of a concgram pass through the statistical thresholds. Yet, in Maisa's list of concgrams, *mother-to-child-transmission* appeared 4 times as MOTHER/TO, CHILD/MOTHER, MOTHER/TRANSMISSION and CHILD/TRANSMISSION.

the texts come from, the normal distribution of words in a text (e.g. Zipf's law)[58] and many others.

Table 5.2 Text coverage by significant concgrams: C1 vs. C2

Student	C1 size	Total instances of significant concgrams, C1	% of words participating in significant concgrams, C1	C2 size	Total instances of significant concgrams, C2	% of words participating in significant concgrams, C2
Hertta	39,449	4,523	23%	123,662	16,909	27%
Kaisa	21,887	3,283	30%	64,809	9,388	29%
Linda	37,641	6,004	32%	133,778	23,340	35%
Maisa	21,501	2,451	23%	85,234	13,864	33%
Nora	14,843	1,414	19%	86,500	11,652	27%
Average			25%			30%

It seems what the comparison presented in Table 5.2 is able to show is that the proportions of text covered by significant concgrams in L2 texts are not remarkable in any way, they are not too small or too large, but just the product of all the factors conspiring in each case. Also, it is clear that the automatically retrieved patterns are only part of what we might assume the total proportion of patterns produced on the idiom principle in the light of previous research which predicts at least 50% of running text or above (Erman and Warren 2000; Altenberg 1998; Dąbrowska 2004; see Section 4.2.2). So I will treat the patterns on the lists as a representative sample: of course these are not comprehensive lists of all multi-word units used by the students in their writing, but unbiased in the sense of being automatically retrieved.

It is probably important to mention that since C2 corpora are corpora of expert writing rather than native-speaker writing, the comparison cannot tell us anything about the possible difference between native and non-native proportions of text comprised of significant concgrams, but rather it might, though bearing in mind the reservations sketched out above, shed light on the difference between an experienced writer of a discourse community and an apprentice.

5.2.2. Comparing C1 patterns to the priming language (C2): Do they match?

So, to what extent are the patterns used by L2 writers the same as what expert writers' in their respective fields use?

[58] The percentages of significant concgrams for C1 and C2 may be similar due to word frequency distributions normal for any text and therefore may not be reflective of the extent to which the idiom principle is available to its authors.

As described in Section 4.2.2, ConcGram allows comparing phraseological patterns of one corpus to the phraseological patterns of another corpus automatically. For this purpose a list of 2-word concgrams from Corpus 1 is compared with Corpus 2. If a certain concgram does not occur in C2 at all, zeros are displayed. Therefore it is possible to say how many of the concgrams from C1 also occur in C2.

Comparing the phraseological patterning of students' writing to the phraseological patterning of expert writers in this automatic way shows that more than half of significant concgrams from students' production data also occur in expert writing of their field of study. Table 5.3 gives the exact figures.

Table 5.3 The percentage of Matching patterns between C1 and C2

Student	Significant concgrams (t≥2 & MI ≥3), C1	Number of C1 significant concgrams (T≥2 & MI ≥3) which also appear in C2	Percentage of C1 significant concgrams (T≥2 & MI ≥3) which also appear in C2
Hertta	464	271	58%
Kaisa	347	194	56%
Linda	524	331	63%
Maisa	260	196	75%
Nora	108	74	69%
Average			64%

The percentages of matching concgrams between the two corpora are higher than might have been expected compared to the earlier research on L2 users' phraseological competence (see Ch. 3). However, without a further point of comparison, it seems difficult to give a more conclusive evaluation. For this reason, I have run several additional comparisons, setting one student's C1 against a different student's C2. The results of this procedure are summarised in Table 5.4. For convenience, the original numbers of matching concgrams between corresponding C1 and C2 are given in the last two columns.

Table 5.4 The percentage of Matching patterns between non-corresponding C1 and C2

C1	C2	Number of C1 significant concgrams (T≥2 & MI ≥3) which also appear in C2	Percentage of C1 significant concgrams (T≥2 & MI ≥3) which also appear in C2	Number of C1 significant concgrams (T≥2 & MI ≥3) which also appear in the **corresponding** C2	Percentage of C1 significant concgrams (T≥2 & MI ≥3) which also appear in the **corresponding** C2
Hertta	Maisa	92	20%	271	58%
Kaisa	Linda	70	20%	194	56%
Linda	Hertta	171	33%	331	63%
Maisa	Nora	74	28%	196	75%

Nora	Kaisa	33	31%	74	69%
Average			26%		64%

It is clear from Table 5.4 that when concgrams from a student's C1 are compared to her own reference corpus, the percentages of matching concgrams are much higher than when they are compared to a different student's C2.

There is another way to get additional insight into these numbers. From the start it was not clear how many of the phraseological patterns actually occurring in C1 were picked up by ConcGram and passed through the statistical thresholds. Since comparing a list of concgrams from C1 against C2 is an automatic process, the statistical thresholds can be manipulated to see how the percentage of matching concgrams in C2 changes accordingly, if at all. The values for t-score and MI value are pre-set in the programme, so in addition to the list of concgrams with both statistical tests applied, concgram lists can be retrieved without any cut-off levels, with only t-score ≥ 2 as a cut-off level, and with only MI value ≥3 as a cut-off level. In the last case, with only MI value ≥3 as a cut-off level applied, it is also necessary to raise the frequency threshold to at least n=3 because of its low-frequency bias, which will be explained below. With very low frequency thresholds, MI tends to bring up many meaningless co-occurrences, such as IS/HORIZON, BUT/LAST and SAMPLE/WELL (see Clear 1993: 280; Evert 2008). The results of these new calculations are presented in Table 5.5.

Table 5.5 The percentage of Matching patterns between C1 and C2 in 3 additional scenarios

Student	C1 concgrams , no cut-off s	Percentage of C1 concgrams with no cut-offs which also appear in C2	C1 concgrams, t≥2	Percentage of C1 concgrams with t≥2 which also appear in C2	C1 concgrams, MI ≥3, **freq. ≥ 3**	Percentage of C1 concgrams with **freq. ≥ 3**, MI ≥3 which also appear in C2
Hertta	19374	59%	986	73%	6302	32%
Kaisa	13479	54%	621	68%	4299	41%
Linda	19537	63%	1100	74%	5276	49%
Maisa	10890	70%	554	84%	2384	62%
Nora	8184	64%	275	75%	2112	51%
Average	14293	62%	707	75%	4075	47%

As can be seen from Table 5.5, when no cut-off points are applied, the percentage of matching concgrams is closest to the one obtained when both statistical cut-off levels are applied. When only t-score is applied, the figure rises, but when only MI value is applied, it drops, even though the frequency threshold is raised to n=3. Let us try to explain the reasons behind this fluctuation briefly.

The t-score measure is known to have a high-frequency bias, that is, when t-score is applied to a list of potential collocations, it tends to prioritise those collocations whose co-occurring words have the highest frequency in a corpus. As a consequence, many so called grammatical collocations get high scores. Therefore, it is not surprising that C1 and C2 are more like each other grammatically or in terms of high-frequency words, than they are lexically, or in terms of low-frequency words.

Mutual information or MI, in contrast, is characterised by a low-frequency bias. It compares the probability of occurrence of the two words forming a concgram/collocation independently with the probability of their co-occurrence, in other words it takes the relative (with respect to the corpus size) frequencies of each of the words and compares them with the relative frequency of their occurrence together. So in order to get a high MI value, two words must either co-occur very frequently or rarely occur independently of each other. To put this another way, if two words are frequent in the corpus over all, they must co-occur frequently as well in order to pass through the MI threshold. If these two-words are low-frequent in the corpus, it is enough that they co-occur just twice to rank high in the list sorted by MI. The good examples of this relationship are concgrams AS/WELL, DASHED/LINES and DE/FACTO from Maisa's C1. *De facto* and *dashed lines* are the most low-frequent concgrams on the list, they only occur twice. However, the words forming them: *de, facto, dashed* and *lines* do not occur at all outside these concgrams. On the list of C1 concgrams retrieved by applying MI threshold of 3, frequency threshold of 2, and sorted by MI value, they rank 6 and 8 respectively. In comparison, the most high-frequent concgram in the corpus AS/WELL occurring 51 times, only ranks 1069 (out of 2384) because *as* occurs 133 times (i.e. 82 times outside the concgram) and is in fact 22nd most frequent word in the corpus and *well* occurs 36 times and is the 98th most frequent word in the corpus. Therefore, it is said that MI tends to pick up many specialised terms, among which we might find topic-specific phrases which may be key to the corpus in question, i.e. they would distinguish between two corpora instead of being shared by them. Accidental co-occurrences may remain a problem even despite the raised frequency threshold.

One more aspect of the obtained results draws attention: the variation in the percentages of Matching concgrams between students in each of the four scenarios. For example, when both statistical tests are applied, the figures vary from 56% to 75%. This variation might be rooted in the students' usage patterns, but this is unlikely. If there are any differences in the way these five students use language, they would be too subtle to be captured by such crude measures. What is more likely is that some students' reference corpora turned out to be better representative of the language which 'primed' them than others', but this is hard to control for. However, we are not interested in the inter-student differences here. The important thing is that apparently L2 users do learn from the language they are exposed to: this is something the percentages are able to confirm. The overlap between C1 and C2 corpora is large in all four scenarios, and the exact proportions of the overlap vary presumably due to the properties of the statistical tests applied, for example the low-frequency bias of MI value and the high-frequency bias of t-score.

5.2.3. How realistic is the automatic comparison? - A qualitative examination

The calculations in Section 5.2.2 show that C1 and C2 corpora overlap to a considerable extent. A natural next step to take is to analyse the lists of concgrams qualitatively. First, this will allow us to check how realistic the automatic comparison was and second, to examine in which ways C1 patterns are similar to C2 patterns or different from them.

By generating and comparing concordance lines for each concgram from both C1 and C2 corpora it is possible to examine whether this concgram indeed represents the same pattern or whether there are differences between corpora. On the basis of such concordance evidence, I have classified concgrams into the following categories. First, a concgram can be *Matching*, in case the pattern it represents in C1 is the same as in C2. Second, it can be *Non-matching,* in case C2 does not exhibit any corresponding pattern to a C1 pattern or the pattern it does exhibit is considerably different from the C1 pattern. Thirdly, it is possible that there is no unit of meaning pattern behind the concgram, in other words the concgram may turn out to be meaningless even though it has passed through the statistical thresholds set. And finally, since I deliberately did not remove proper names from any of the two types of corpora, some C1 concgrams can be formed by one or even two proper names. These concgrams are valid as being characteristic of C1 but are not expected to occur in C2, and are excluded from the analysis. Concgrams which include at least one proper name are classified as *Names*. Table 5.6 shows the numbers for these four categories calculated for two data sets. The table also presents the total number of significant concgrams in C1, excluding concgrams which are formed by proper names and concgrams which do not represent units of meaning.

Table 5.6 Qualitative comparison and classification of C1 concgrams against C2

Student	Matching concgrams	Non-matching concgrams	Total units in C1	% of matching concgrams	Names	Not units

Kaisa	138	81	219	63%	20	139
Maisa	153	48	201	76%	27	32

As can be seen from the table, when the data is analysed manually, the percentage of matching concgrams remains high.

I will now examine Matching and Non-matching groups of concgrams in more detail using these two data sets. In the case of Matching patterns it is interesting to explore what kind of patterns are 'acquirable' from exposure and how closely L2 writers follow the patterning they were exposed to. In the case of Non-matching patterns the intriguing question is why they do not match C2 patterning.

5.3. Matching patterns

As mentioned in Sections 4.2.3 and 5.2.3, all the significant concgrams from two C1 corpora, Maisa's and Kaisa's, are analysed using concordance data from both C1 and C2. The complete lists of Matching and Non-matching concgrams for the two students are presented in Appendix B. For reasons of space, it is not possible to provide a full analysis of all the Matching and Non-matching patterns. Only a limited number of examples will be presented, all selected on the basis of the preliminary qualitative analysis of concordances for all the extracted concgrams. The reasons behind the selections of Matching and Non-matching patterns are different because their respective analyses are intended to shed light on different questions. In the case of Matching concgrams, it is interesting to know how closely C1 patterns can match C2 patterning, in the case of Non-matching – why they do not match C2 patterning.

Most of Matching patterns are very conventional and do not tell us anything new about L2 language patterning or learning from exposure. There are many grammatically dictated co-occurrences such as *does not* or *have been,* common lexical patterns such as *for instance, focus on, based on, so far, relationship between*, *even though, carried out*, or *on the other hand* and technical terms forming simple collocations of two words such as *sentinel survey, uncertainty analysis, life expectancy* or *growth rate* typical of Maisa, a demographer, and *comparative method, false friends, language family, orthographic similarity* or *morphological analysis* typical of Kaisa, a computational linguist. Such patterns have already been analysed many times in different studies. Also, the analysis of their usage patterns with the help of concordances has not revealed anything unexpected. Thus, they are not taken up for further analysis. Instead, we would be interested to find extended patterns which are unlikely to have been acquired explicitly through formal classroom instruction or look-up in the dictionaries, but rather are likely to have been implicitly learned from exposure.

So the primary goal of this chapter is to find out to what extent C1 patterns match C2 patterning not only quantitatively, but also qualitatively. In particular, we will look at how common language patterns specialise in field-specific use in Section 5.3.1 and the level of detail to which C1 patterns can match C2 patterning in Section 5.3.2. Section 5.3.3 already prepares us for the analysis of the Non-matching patterns. It shows how a preference may develop for a certain pattern and as a result make it more frequent than in the source texts, or how the difference in frequencies of a pattern between C1 and C2 can be determined by content and thus reflect the "aboutness" of text (Scott and Tribble 2006) rather than illuminate any user-specific preferences. Later we will see similar patterns among Non-matching concgrams.

As pointed out in Section 4.2.5, sometimes the BNC will also be used to complement the comparative analysis of C1 and C2 patterning. It is important to repeat here that in case C1 pattering matches the BNC instead of C2 or in addition to C2, this does not undermine the argument of L2s learning holistically from exposure rather than constructing the pattern on their own.

5.3.1. Specialisation of patterning

This section will give two examples of patterns from the Matching-category which seem to be developing field-specific associations.[59]

The concgram ASSUMPTIONS/ABOUT (13/27) from Maisa's data set presents a case in point. It gives the impression of a general use pattern because there is nothing field-specific in the collocation itself, and indeed *about* is the most significant (determined by the log-likelihood value) collocate of *assumptions* in the BNC too. However, in the two corpora one usually makes *assumptions about* certain demographic rates, e.g. *fertility, mortality*, a future trend/change. That is, a general pattern becomes specialised, even 'terminologised', so to speak. In addition to this, the verb MAKE seems to be participating in the pattern as a collocation, which is a frequent pattern in the BNC too. Examples below show selected concordance lines for the concgram ASSUMPTIONS/ABOUT generated from Maisa's C1 (5.7) and C2 (5.8):

[59] Cf. Sinclair on specialisation of meaning (e.g. Sinclair 2004: 30, Sinclair et al. 2004: xxii).

(5.7)

1 so they need to be chosen carefully. 1.1 HIV **Assumptions about HIV** need to be **made** first,
2 latest year where TFR is available is 2006–07, **assumptions about fertility** need to be made from
3 in the mortality **assumptions**, but not in the **assumptions about fertility**. In addition,
4 Five different scenarios are produced using **assumptions about** different levels of behavioural **change**
7 the method requires that its user **makes assumptions about future trends**, but does not
9 five different scenarios are produced using **assumptions about** different levels of behavioural **change**
10 The next three, more likely scenarios rely on **assumptions about future** behavioural **changes**,
11 studies will also be used in order to **make assumptions about** current levels of **fertility**,

(Maisa, C1, concordances for the concgram ASSUMPTIONS/ABOUT)

(5.8)

1 determinants and correlates in preparing **assumptions about future fertility trends**. Logistic
2 be carried to term. Spectrum allows us to modify **assumptions about** the **fertility** of HIV positive
7 data evaluation, parameter2 estimation, **making assumptions about future change**, and final
8 by applying the cohort-component method to **assumptions about fertility, mortality**, and
9 the current demographic situation in countries, **assumptions about** the **future** can turn out to be
10 of projection inputs: parameter estimates and **assumptions about future change**. An essential
11 can be used to **make** reasonable and consistent **assumptions about** the **future** course of **fertility**,
13 Bureau uses models that take into account **assumptions about mortality** and **population size**

(Maisa, C2, concordances for the concgram ASSUMPTIONS/ABOUT)

One possible observation based on this example is that when a unit begins to be used for a special purpose, its pattern develops: it attracts new components and the components it already had become more fixed, i.e. they shift from being more abstract associations like a semantic preference or a colligation to collocating verbatim. The new or field-specific components, of the *assumptions about* unit, like a possible semantic preference for demographic constructs and variables, are reproduced in Maisa's writing.

Table 5.7 presents some of other similar cases from Maisa's and Kaisa's use.

Table 5.7 Examples of patterns which have become specialised in Maisa's and Kaisa's C1

Co-occurring word	Co-occurring word	C1	C2	Pattern represented by the concgram
Maisa				
HIGHER	THAN	11	33	*higher than* (used to talk about prevalence, incidence, number of something)
LIKELY	MORE	6	28	*be* + *more likely* + *to*-inf. (used to compare different groups of people and their properties)
DURING	PERIOD	7	16	*during* (e.g. *the survey* [C2]/*projection* [C1]) *period*
INFECTIONS	NUMBER	10	19	*(the) number of new/HIV/new HIV infections*
INFECTIONS	NEW	18	51	

LIVING	PEOPLE	5	31	(*number/percentage/amount of) people living with HIV*
AGE	SPECIFIC	6	23	*age-specific (mortality rates, fertility rates, prevalence)*
AMONG	WOMEN	8	124	*among (young/er, married, pregnant* etc.) *women*
AVAILABLE	DATA	23	57	*available (HIV prevalence, surveillance) data; data (is/are/were /becomes,* etc.*) available*
BIRTHS	NUMBER	11	6	*number of (live/total) births*
OVER	TIME	7	67	*over time*: e.g. CHANGE/*trends* **** *over time*
Kaisa				
AT	LEVEL	4	15	*at the (feature, character, sentence, proto-language) level*
ON	SIDE	12	6	*on the (left, left/right-hand, theoretical, lexical, surface) side*
ARE	PERCEIVED	6	3	*are perceived as similar*
SEVERAL	THERE	8	2	*There are/have been several* + pl. noun (*approaches/attempts/ways* etc.)

Another pattern I would like to discuss here is of a different kind, but it also shows the possible impact of specific priming on language use. The pattern is represented by two matching concgrams: CLOSELY/RELATED (11/9) and LANGUAGES/RELATED (16/17). It does not come as a surprise that the two concgrams overlap in a combined 3-word concgram CLOSELY/RELATED/LANGUAGES and also in a fixed trigram *closely related languages* which seems to be a usual final stage when the pattern is developing (see Section 5.5.2). At the same time the overlap is not total, and it is interesting to investigate the divergent points as they are likely to throw light on the eventual process of convergence. In order to get an idea of the distribution of the pattern among the texts in C1, I will start by generating concordances with AntConc, since it allows making searches in a group of text files instead of one single merged file. The query is made for *related languages* which is 'superordinate' in relation to *closely related languages.* The concordance lines from C1 (5.9) show that the co-occurrence is spread out across 7 out of 12 of Kaisa's drafts:

(5.9)

1	ple cognate lists of **closely related languages**, are used to refine NLP appl	kaisa_4_100809.txt
2	hods align relatively **closely related languages**, English and French data, so	kaisa_10_300410.txt
3	to use cognates of **closely related languages** to aid in machine translation	kaisa_10_300410.txt
4	performs best on **closely related languages**. With more remotely related	kaisa_10_300410.txt
5	forms best on **closely related languages**. With more remotely relate	kaisa_critique_110209.txt
6	come from two **distantly related languages**, choosing the cognates that	kaisa_3_250609.txt
7	ge. In the case of **distantly related languages**, like Finnish and Sumerian all	kaisa_9_300410.txt
8	guage. These **genetically related languages** are considered to have been di	kaisa_2_140509.txt
9	ge from which **genetically related languages** descend from. Languages are	kaisa_2_140509.txt
10	ages. With more **remotely related languages**, and with languages with hea	kaisa_10_300410.txt

11 s. With more **remotely related languages**, and with languages with heavy kaisa_critique_110209.txt
12 that take place when two **related languages** separate from their proto-lan kaisa_9_300410.txt

(Kaisa, C1, concordances for *related languages*)

The remaining RELATED/LANGUAGES co-occurrences with intervening words in between and therefore not captured by AntConc but readily seen in ConcGram are given in Example (5.10):

(5.10)

1 distance. He points out that even if the **languages** are **related**, the task of proving the
2 is successful or not. Sister language: **languages** which are closely **related** to each other
3 with reconstructing the proto-language of **languages** that are less closely **related** to each
16 that Sumerian could be **related** to Uralic **languages** was ever since ruled out by both

(Kaisa, C1, instances of the concgram RELATED/LANGUAGES exhibiting positional and constituency variation)

Now we see that *closely related languages,* occurring 5 times in 3 different drafts, is the dominant pattern. But there are also *distantly related languages* (n=2), *genetically related languages* (n=2) and *remotely related languages* (n=2). Let us now compare these patterns to the BNC patterning.

A query in the BNC shows that the 5 most significant (by log-likelihood value) adverbs collocating with *related* are: *closely, directly, distantly, intimately, inversely* (in the order of significance), so both *closely* and *distantly* are at the very top. *Genetically* is 22nd, not surprisingly because it is much more specific (a hyponym rather than a hyperonym like *closely*). As for *remotely*, it occurs only 3 times in the whole BNC which shows that this is a very rare pattern which is unlikely to be encountered in general use. So is it an acquired or a constructed pattern in Kaisa's case? Let us now look at the patterning of Kaisa's reference corpus in Example (5.11).

(5.11)[60]

1 problem is even more acute for **closely related languages** that have different stress rules. 8 In
2 **languages**, there are other **closely related languages** that are retained in the training set,
3 appropriate when dealing with **closely related languages** (e.g., Dutch and German), which share a
4 most **closely related** among the four Algonquian **languages**, according to all measures of phonetic
5 the accuracy on genetic cognates. However, for **languages** that are unrelated or only **remotely related**,
6 Genetic Cognates are word pairs in **related languages** that derive directly from the same word
7 of an item from its similarity in **related languages** goes back a long way in the
8 linguistics, cognates are words in **related languages** that have developed from the same

[60] If not otherwise stated, all hit lines are included in the concordances presented which sometimes contain irrelevant lines: for convenience, such lines are struck through like in Example 5.11, line 16.

9 identify a large portion of cognates in **related languages** without explicit knowledge of
10 unordered wordlists from two or more **related languages**, and produce a list of aligned cognate
11 and align cognates in vocabularies of **related languages**. Nevertheless, thanks to its grounding
12 directly in the vocabularies of **related languages** by combining the phonetic similarity of
13 cognates in the vocabularies of **related languages**. I show that a measure of phonetic s
14 and align cognates in vocabularies of **related languages** (e.g. colour and couleur). Nevertheless,
15 However, in the case of very **remotely related languages**, the difference may no longer be
~~16 same language family. For each of the removed **languages**, there are other closely **related**~~
17 is particularly interesting, because these two **languages** are not **closely related**. Some of the
18 set of 82 cognate pairs from various **related languages**. The distance function is very simple;

(Kaisa, C2, concordances for the concgram RELATED/LANGUAGES)

As can be seen, Kaisa's C2 reveals 2 occurrences of REMOTELY/RELATED/LANGUAGES. Both instances come from the articles Kaisa used for writing a summary and a critique later incorporated into the thesis itself. The first time Kaisa used the pattern is in the critique, from which the use travelled to the main body of the thesis. So it is possible that the use in the two articles she read primed her own use.

As for other adverbs collocating with RELATED/LANGUAGES in Kaisa's C2, we also see 5 occurrences of CLOSELY/RELATED/LANGUAGES, out of which 3 form a fixed trigram. Yet, there are no occurrences of DISTANTLY/RELATED/LANGUAGES or GENETICALLY/ RELATED/LANGUAGES: these two usages must have come from somewhere else. In any case it is clear that these two are not Kaisa's own idiosyncratic phrasings because they are frequent in the BNC.

The examples given in this section show how some patterns while being at their core common in general language use reveal signs of forming field-specific associations. The collocation *assumptions about* seems to be developing a semantic preference for demographic rates, e.g. *fertility, mortality.* The collocation *related languages* associates not only with the adverbs frequent in the BNC, like *closely* and *distantly,* but also with an adverb *remotely,* which is very rare in the BNC but occurred twice in Kaisa's priming corpus. This evidence suggests that the students have acquired certain subtle patterning that is traceable to the texts they were exposed to.

5.3.2. How nuanced can matching be?

An example of a remarkably exact matching is the patterning of the concgram TAKEN/INTO coming from Maisa's data set. The concgram occurs 14 times in Maisa's C1 and 7 times in her C2. And again in (5.12) and (5.13), I am only giving some selected concordance lines because they are all similar.

(5.12)

1 reduction due to HIV infection is **taken into account** in the default patterns, according to
2 2 The research setting The three factors **taken into consideration** in a population projection are
3 addition, international migration was not **taken into account** in this projection because its
4 in the projection. National migration was **taken into account** on these regional projections
…
13 [ref.] These errors can be **taken into consideration** in the projection as they are w
14 scenarios. HIV AIDS is naturally **taken into account** in the mortality assumptions, but not

(Maisa, C1, concordances for the concgram TAKEN/INTO)

(5.13)

1 epidemic. The needs of end users should be **taken into account** when building up second generation
2 epidemic and gender differences must be **taken into consideration**. The WHO Regional Office for

(Maisa, C2, concordances for the concgram TAKEN/INTO)

The interesting aspect of the matching pattern is that in both corpora there are only two words which complete it: *account* and *consideration,* with *account* being the preferred collocate. Extending the search and looking for all the forms of the lemma TAKE co-occurring with the preposition *into* reveals that in C1 *consideration* completes the pattern 4 times and *account* 16 times, i.e. 80% of the time, in C2 *consideration* completes the pattern 3 times and *account* 13 times, i.e. 81% of the time.

The BNC yields very similar results. Querying for the pattern "{take} **** into", which equals the search for the concgram TAKE*/INTO in the internal span of 4 in our C1 and C2 corpora, and looking at the list of its collocates of up to three to the right show that indeed *account* is the most significant collocate which is also much more frequent that the second most significant collocate, *consideration*. Together they are used 3144 times in the pattern in question, with *account* appearing 90% of the time. As such, in all the three corpora, C1, C2 and the BNC, type-token frequency distribution[61] of the items comprising the category of collocation[62] for the unit TAKE*/INTO is very similar (see Table 5.8).

[61] Cf. Ellis et al.'s work on type-token frequency distribution of verbs in verb argument constructions (e.g. Ellis and Ferreira-Junior 2009; Ellis et al. 2013; Ellis et al. 2014).

[62] Since there is more than one item which can realise the category and the items form a semantic set, the category can be interpreted as a semantic preference. However, since the association formed between each of the two items and the unit is quite clearly verbatim, and word association data confirm this too, I prefer to analyse them as alternating items in the collocation category. If, for example, a Russian ESL user produces TAKE*/INTO + *attention*, perhaps under the influence of a similar expression in Russian *принять во внимание*, I would say that the category of collocation has been approximated in this case to semantic preference to include *attention* on a par with *account* and *consideration.*

Table 5.8 Type-token frequency distribution of *account* and *consideration* in the unit TAKE*/INTO in C1, C2 and the BNC

	C1		C2		BNC	
	N	coverage	N	coverage	N	coverage
[1]TAKE*/INTO + collocation	20	100%	16	100%	3144	100%
TAKE*/INTO + *account*	16	80%	13	81%	2839	90%
TAKE*/INTO + *consideration*	4	20%	3	19%	305	10%
[1]with the meaning (i.e. semantic prosody): "consider something along with other factors before reaching a decision" (*Oxford Dictionary of English* 2010).						

All the three corpora show a variation between the two nouns *account* and *consideration* and an overwhelming preference for *account*. So this time not only the collocational patterning of the unit is respected but also the distributional asymmetry of the collocates is replicated, reflecting the language users' ability to acquire "statistical knowledge": their "sensitivity to frequency" and ensuing implicit "tallying" (Ellis 2002, 2006, 2009, 2012a).

Let us now look at the next example (5.14). The concgram MORE/RAPIDLY (5/1) coming from Maisa's data set is not particularly impressive at first sight. It is a Matching one because it occurs in C2 too (5.15), even if just once.

(5.14)

1 HIV incidence, however, seems to be diminishing **more rapidly**. For the population projection, two
2 with the treatment. Urban population is growing **more rapidly** in size than rural population, as was
3 percentage. Incidence, however, is declining **more rapidly**, which means that the amount of new
4 In the projections, urban population is growing **more rapidly** than rural, as expected. If there we
~~5 falling very **rapidly**, probably indicating that **more** people are staying alive because of ART and t~~

(Maisa, C1, concordances for the concgram MORE/RAPIDLY)

(5.15)

1 between groups. Because behaviour changes **more rapidly** among young people than among older

(Maisa, C2, concordances for the concgram MORE/RAPIDLY)

In examining the concgram MORE/RAPIDLY, it seems reasonable to take into account the usage of the lemma RAPID as a whole. However, just like in the previous example with the concgram TAKEN/INTO, where the form *taken* was actually representative of the usage of the lemma TAKE as a whole (it occurred 14 times in the form *taken* and only 6 in other forms), MORE/RAPIDLY turns out to be representative of the usage of the lemma RAPID as the concordances for “rapid*” (5.16) generated with AntConc show.

(5.16)

1 same trend. The growth rate has been **declining rapidly** in 1991–2001 and will maisa_8_010311.txt
2 of epidemic types including slowly growing, **rapidly growing** and stable epidemics. maisa_3_140110.txt
3 Between 2003 and 2007, ART coverage **increased rapidly** from 3 per cent to 60 maisa_5_031110.txt
4 Incidence, however, is **declining more rapidly**, which means that the amount of maisa_8_010311.txt
5 incidence, however, seems to be **diminishing more rapidly**. For the population maisa_8_010311.txt
6 urban population is **growing more rapidly** than rural, as expected. If there were maisa_7_280211.txt
7 Urban population is **growing more rapidly** in size than rural population, maisa_8_010311.txt
8 and the first birth occurring **rapidly**, approximately in the two following years. maisa_1_151009.txt
9 age of people living with HIV is not **falling** very **rapidly**, probably indicating that maisa_8_010311.txt

(Maisa, C1, concordances for the lemma RAPID)

First, RAPID is used only as an adverb. Second, *more* is its frequent collocate as it appears together with *rapidly* 4 times out of 9 instances. Third, the emerging unit has a semantic preference for verbs with a sense of changing up or down the scale: *growing, increasing*; *declining, diminishing, falling.* Fourth, the unit colligates with the Progressive aspect. So the concgram MORE/RAPIDLY proves to be a surface feature which readily lends itself to phraseological tools such as ConcGram but has a more complex pattern of an extended unit of meaning. The growing number of such cases suggests that it must have some connection to the mechanism of unit formation. It seems that the gradually developing pattern, the one acquiring semantic and structural associations, is likely to give rise to a verbatim association, i.e. a collocation, in the end.

Returning to the matching features of Maisa’s pattern, it seems useful to search for the concordances of the adverb *rapidly* in C2 (5.17) in addition to one concordance line for the concgram MORE/RAPIDLY.

(5.17)

1 epidemic types including slow growing epidemics, **rapidly growing** epidemics, and stable epidemics in
2 The study showed that HSV-2 prevalence **increases rapidly** with age. By the age of 18 years, 60% of
3 sub- populations. In Eastern Europe, for example, **rapidly changing** social circumstances had led by
4 lation. Concentrated • Principle: HIV has **spread rapidly** in a defined sub-population, but is not
5 prevention of mother to child services have been **rapidly rolled out** in Namibia. These services
6 ners and service providers with the challenge of **rapidly** scaling up institutional and community
7 ulation). Concentrated Epidemic ? HIV has **spread rapidly** in at least one defined sub-population,
8 to avoid this scenario if prevention efforts are **rapidly rolled out**. Otherwise the continuing
9 on between groups. Because behaviour **changes** more **rapidly** among young people than among older
10 as reported in Namibia in 1986. The epidemic **grew rapidly** in the 1990s until 2002, apparently
11 is by no means inevitable, an epidemic can **shift rapidly** between one state and another, and the
12 of as a priority. Even when HIV prevalence **rises rapidly** in defined sub-populations, countries may
13 ehaviour is relatively low is unlikely to **change rapidly**. It is therefore recommended that such

14 r for children under 3 years of age and **decrease rapidly** for children who have survived past those

(Maisa, C2, concordances for the word *rapidly*)

The concordances in (5.17) show that in contrast to C1 patterning, *rapidly* in C2 does not consistently collocate with *more* or any other word and does not have an association with the Progressive aspect. Yet there is a semantic preference for 'change' verbs like INCREASE, CHANGE, SHIFT, SPREAD, GROW, DECREASE, though the 'change' they designate is not necessarily happening up or down the scale. In comparison to this pattern, C1 patterning is more fixed which can be regarded as an example of the process I call 'fixing' (see Section 5.5.2).

What we have so far is an association of the adverb *rapidly* with verbs describing some kind of change. The question is whether this is a specific function of the adverb *rapidly* or whether there are other adverbs which are used for the same purpose. To shed light on this issue, I examined concordance lines in Maisa's C1 and C2 for all possible adverbs with a similar meaning: *fast, quickly, swiftly, promptly* and *speedily*. In all the concordance lines, I have found just one usage in Maisa's C2 (5.18) which is similar to the pattern in which the adverb *rapidly* occurs.

(5.18)

as. But in most generalized epidemics, infection **quick**ly **spreads** from urban areas along major transport r

(Maisa, C2)

There are not many uses of adverbs with this meaning altogether (5.19). And all of them come from Maisa's reference corpus (C2), while Maisa herself only uses the adverb *rapidly*.

(5.19)

1 . When first marriage is universal and early, and **quickly** follows the onset of the fertile period, as was
2 marriage was 21·5 years, and first birth followed **quickly** within two years (age 23·6 on the average). The
3 ic, EPP will often allow the curves to grow very **quickly** at the start of the epidemic. This can be constra
4 safe behaviour is therefore reflected much more **quickly** in lower STI rates than it is in lower HIV rates.
5 as. But in most generalized epidemics, infection **quickly** spreads from urban areas along major transport ro
6 ulations required – computer programs handle this **quickly** and painlessly – but, rather, the derivation of
7 up, the Ovambo, were in a special situation, with **fast** increasing age at marriage and average level of

(Maisa, C2)

The function of describing the speed with which something is changing has been given entirely to the adverb *rapidly* in both corpora.

Finally, let us compare the patterning of the adverb *rapidly* in C1 and C2 with the BNC patterning. Briefly, the first 10 most significant collocates of the adverb in the span 3 to the right and 3 to the left are*: growing, changing, expanding, rising, grew, increasing, expanded, more, very* and *becoming*. So here too, there seems to be a strong association with verbs meaning some kind of 'change' and, grammatically, with the *-ing* form of the verbs. Yet, a quick look at the concordance lines of the first three collocating verbs shows that they mostly play the role of an adjective modifying a noun rather than a predicate in the Progressive aspect, like in *a rapidly changing world* or *a rapidly growing number.* In any case, the BNC patterning supports the found semantic preference for 'change', some of the specific collocating verbs are also the same: the lemmas GROW and INCREASE are frequent in all three corpora.

In contrast, information about the semantic preference, grammatical associations (colligations) and specific collocations is not provided in the *Oxford Dictionary of English* (2010), for example. The definition given there is: "very quickly; at a great rate". Even though the examples provided are very similar to the frequent patterning of the BNC: "*the business is expanding rapidly; the problem is rapidly worsening"*, the patterns which are common to the adverb are not explicitly stated. This gives reason to suppose that Maisa has acquired the patterning she displayed in her writing from exposure rather than through explicit learning.

In sum, we can conclude that in both cases, in the units of meaning TAKE *into account/consideration* and 'changing' *rapidly*, Maisa's patterning follows expert writers' patterning or the patterning of general language use remarkably closely, both in terms of the level of detail and in terms of frequency distributions. This is unlikely to have been learned through classroom instruction or language reference materials, and therefore can be argued to have been acquired through exposure, on the idiom principle.

5.3.3. Matching but 'overused' patterns

Several concgrams stand out from the list of matching concgrams because they are much more frequent in C1 than in C2, despite adjusting for size. Here I will analyse some Matching concgrams with the largest differences in frequencies:

The first example comes from Maisa's data set, it is a concgram ANTIRETROVIRAL/TREATMENT appearing 14 times in her own writing but just 3 times in the expert writing. In fact, concordance lines in (5.20) show that the concgram occurs in C2 only twice.

(5.20)

1 Indicator Survey ANC antenatal care (clinic) ART **antiretroviral treatment** ARV **antiretroviral** (drug)
2 one below. Here you can describe the scope of **antiretroviral treatment**. A. Proportion surviving
~~3 care (clinic) ART **antiretroviral treatment** ARV **antiretroviral** (drug) BSS behavioural~~

(Maisa, C2, concordances for the concgram ANTIRETROVIRAL/TREATMENT)

But as concordance lines in (5.21) show, in C1 it is truly frequent.

(5.21)

1 also as input data. Next, ART (**antiretroviral treatment**) coverage data are needed, as in number
2 tical scenario, there would be no **antiretroviral treatment** and the HIV AIDS situation would be
3 ates of individuals needing ART (**antiretroviral treatment**) are also crucial when planning
4 asidence, and whether they were on **antiretroviral treatment** (ART). (Ibid., 10; 33.) The 2008
5 ults. Also, the consequences of **antiretroviral treatment** can be detected here. If there were no
6 of the results, the initiation of **antiretroviral treatment** has brought a visible change in many
7 ould decline more if there were no **antiretroviral treatment**. The amount of deaths, on the
8 ery possible here. If there were no **antiretroviral treatment**, there would be even less mothers
9 howing the impacts if there were no **antiretroviral treatment**. In this scenario, the HIV prevalence
10 e slightly bigger, 2.61 million. If **antiretroviral treatment** had never existed, the population
11 s of this projection. The effects of **antiretroviral treatment** are also evident in sub-Saharan
12 tries. More data is needed regarding **antiretroviral treatment** in order to project better future
13 dynamics but the growing coverage of **antiretroviral treatment** is diminishing the consequences of
14 owth. The impacts of HIV treatment, **antiretroviral treatment** (ART), will be assessed in the

(Maisa, C1, concordances for the concgram ANTIRETROVIRAL/TREATMENT)

Antiretroviral treatment seems to be a term which is important for the topics discussed both in C1 and C2, and its infrequency in C2 raises suspicion. Concordances for *antiretroviral* generated from C2 (5.22) reveal that *antiretroviral* also co-occurs with *therapy*, which in fact constitutes a preferred word combination in C2.

(5.22)

1 searching for alternatives because **antiretroviral therap**y is now altering the natural history of
2 ed. These range from short- course **antiretroviral therap**y during pregnancy to the avoidance of
3 officers. In this era of scaling up **antiretroviral therap**y (ART), a growing number of countries
4 death from AIDS in the absence of **antiretroviral therap**y (ART) are available for developing4 and
5 Clinic ARV **Antiretroviral** ART **Antiretroviral Therap**y AZT 3TC Zidovudine and lamivudine (anti
6 s the number of people in need of **antiretroviral therap**y (ART), the number of orphaned children,
7 all years (past and future). 3.2.4 **Antiretroviral Therap**y in Adults The number or percent of
8 ions: AIM, AIDS Impact Model; ART, **antiretroviral therap**y; EPP, Estimation and Projection Package
9 s in the IDB include the impact of **antiretroviral therap**y (ART) for selected countries. The
~~10 Syndrome ANC Antenatal Clinic ARV **Antiretroviral** ART **Antiretroviral Therap**y AZT 3TC Zidovudine~~

(Maisa, C2, concordances for the word *antiretroviral*)

Judging by Internet sources and concordance lines in (5.22), formally ART stands for *antiretroviral therapy* rather than *treatment*, but *antiretroviral treatment* is indeed sometimes used. For Maisa this has become a preferred alternative, moreover, she does not use *antiretroviral therapy* at all.

Examples of preferred alternatives are not only lexical, there are also grammatical ones. One of them is revealed by two concgrams from Maisa's data set: ASSUMED/IT (21/3) and ASSUMED/THAT (20/3) which overlap just like the concgrams CLOSELY/RELATED and LANGUAGES/RELATED did. Example (5.23) shows the concordance lines for the concgram ASSUMED/IT:

(5.23)

1 coverage versus rural ART coverage. Here, **it** is **assumed that** the ratio is 70% for urban and 30%
2 begin to use medication in 2006 (**it** is **assumed that** 30% of the people cannot tolerate the
3 the end of the projection period. Here, **it** is **assumed that** the same development will continue;
4 (2010). For HIV infected women aged 15–19 **it** is **assumed that** they are sexually active and thus e
5 which I will use in my projection. **It** is **assumed that** during the projection period, the
6 are also affected by education. **It** is **assumed that** education improves people's health,
7 It also changes the fertility levels, and **it** is **assumed that** the fertility of HIV positive women
8 population scenario called "Continuation" **it** is **assumed**, **that** if current (here: 1994–1998)
9 account the different educational groups. **It** is **assumed that** those with higher education are more
10 development for the life expectancy; **it** is **assumed** to stay the same for both males and
11 population) were not included. **It** is **assumed that** these groups are relatively small
12 Bureau of Statistics (GRN NPC 2001), **it** is **assumed that** international migration is not on
13 were living in urban areas in 2001, and **it** is **assumed that** this proportion will grow to 43% in,
14 the need for ART is still significant. **It** was **assumed** in the research setting, **that** the young
15 to 4.2 in 2000 and to 3.6 in 2006–07. **It** was **assumed** to decline further to 2.6 in 2020. The
16 mention which data he is using, but **it** can be **assumed that it** is the 1991 census. The article is
17 the levels of mortality considerably. **It** can be **assumed**, **that** any gains from primary health care
18 the base data is from the 1991 census, **it** can be **assumed that** the division between urban rural is
19 censuses fit well into this pattern. **It** can be **assumed**, **that** the fertility decline is still.
20 the same age. For this age group **it** is generally **assumed that** fertility ratio is 1.2 when compared
21 for Namibia. Even though **it** has been previously **assumed that** HIV AIDS will dramatically decrease

(Maisa, C1, concordances for the concgram ASSUMED/IT)

As can be seen, except for lines 10 and 15 which have subject-to-subject raising (see e.g. Carter and McCarthy 2006: 789-790): 'something *is assumed to* do something', the structure can be presented as follows: anticipatory *it* as subject + mental process verb in the passive + *that*-clause as direct object.

In contrast, this structure is used in C2 only twice as the concordances in (5.24) show, with the third instance representing a different pattern:

(5.24)

1 clear that HIV was a global phenomenon, **it** was **assumed** that the epidemic would follow roughly the
2 levels of infection have been found **it** has been **assumed** that the epidemic is still at an early
~~3 and sex throughout its lifetime, exposing it to assumed age- and sex-specific mortality,~~

(Maisa, C2, concordances for IT/ASSUMED)

The verb ASSUME itself is quite frequent in C2, but used in different patterns. It occurs not only in sense 1: “suppose to be the case, without proof” as in C1, but also in the sense “take or begin to have” (*Oxford Dictionary of English* 2010), which is not relevant for the comparison.[63] In the first sense, the verb ASSUME occurs in the active voice (n=8), as in (5.25); as an adjective or a participle (n=11) as in (5.26); and in a subject-to-subject raising structure which overrides the anticipatory *it*-structure (n=14) as in (5.27).

(5.25)

23 evidence is available in country. Default values **assume** fertility in HIV positive women ages 15-19
25 e rate of progression from infection to death: We **assume** HIV infected people are subject to the

(Maisa, C2)

(5.26)

2 and sex throughout its lifetime, exposing it to **assumed** age- and sex-specific mortality, fertilit
3 on is based on recent data, projected targets, or **assumed** levels. If the last observed or target
26 and other sources. ART coverage is projected by **assuming** a constant yearly percent reduction in
27 . Single, dual, and triple therapy are projected **assuming** a phase out of single-dose and dual

(Maisa, C2)

(5.27)

5 ancy or HIV infection) and their HIV disease is **assumed** not to have progressed to a point where
10 evel until 2012. Coverage of second line ART was **assumed** to be 50 percent in 1998, remaining at

(Maisa, C2)

However, in C1 the verb ASSUME (which occurs in its first sense only) occurs in all the same structures too: in a subject-to-subject raising structure (n=8) as in (5.28); as a participle (n=1) in (5.29); in the active voice (n=4) as in (5.30).

(5.28)

2 ily to 114 500 by 2013. The number of children is **assumed** to grow and then to stabilize at around 7
18 e as nationally. The impact of AIDS on deaths is **assumed** to be heaviest by 2011; after that the

(Maisa, C1)

[63] In accordance with the principle, discussed in Chapter 2, a different sense associates with a different formal representation, i.e. there is nothing surprising in the fact that the same verb in its different sense has a different patterning.

(5.29)

28 to years 2016–2020. Another scenario is made by **assuming** a declining HIV prevalence in 2016–2020.

(Maisa, C1)

(5.30)

29 until year 2015. In 2016–2020, the first scenario **assumes** that HIV prevalence is staying constant
30 n from EPP is imported to Spectrum, the programme **assumes** that the HIV prevalence will stay at the

(Maisa, C1)

So there is no evidence to suggest that Maisa does not know other structures or feels uncomfortable with them. The difference is only between Maisa's preference for using ASSUME in anticipatory *it*-construction and expert writers' for using it in a subject-to-subject raising structure instead. One possible reason for this can lie in the relative fixedness of the anticipatory *it*-construction. In Maisa's case it is realised in an almost completely fixed form: *it is/can be assumed that* which serves her as a convenient opening of a sentence.

In addition to developing individual preferences like in the cases of *antiretroviral treatment* vs. *antiretroviral therapy* and an anticipatory *it*-construction, there are also cases of 'overuse' which seem to reveal meanings and concepts important for the text, rather than usage habits of the writers. One example of such a case is Maisa's concgram DETERMINANTS/PROXIMATE (19/2), examples of which are given in (5.31).

(5.31)

1 research by using the concept of **proximate determinants** of fertility. This concept and its
…
18 (1984) have applied the analysis of **proximate determinants** to the research of sub-Saharan
19 can be detected, but it is the **proximate determinants** that need to be studied in order to

(Maisa, C1, examples of concordance lines for the concgram DETERMINANTS/PROXIMATE)

Proximate determinants (of fertility) is a valid sociological term (see e.g. Poston and Bouvier 2010) which is also used in C2, but only twice (5.32):

(5.32)

1 health and program coverage, and the **proximate determinants** of fertility. Trends in women's
2 [ref.]. "Revising the **Proximate Determinants** of Fertility Framework: What have we

(Maisa, C2, concordances for the concgram DETERMINANTS/PROXIMATE)

Thus, the reason for its ‘overuse’ or considerably higher frequency in C1 in comparison to C2 may be explained by its importance or keyness to C1 texts, i.e. Maisa’s Master’s thesis. This is what she writes about and therefore has to use rather than prefers to use as an alternative to some other wording. Keyness is a textual feature rather than one related to acquisition or processing. As Scott and Tribble (2006) put it, it is “what the text ‘boils down to’” in terms of its propositional content and what makes it different from other texts in terms of the information it communicates. Keyness of text can be revealed through key words which reflect this keyness “avoiding trivia and insignificant detail” (Scott and Tribble 2006: 56). But keyness is visible through larger units too. For example, Warren (2010) suggested the term “aboutgram” for key concgrams, i.e. concgrams which just like key words represent the aboutness of a text.

I did not aim at finding aboutgrams in students’ texts. However, due to the kind of methodology I used, comparison of one text, a Master’s thesis, to other texts in the same field, some aboutgrams could have been retrieved as by-products. Identification of keywords or key phrases on the basis of a comparison of one text with a reference corpus or corpora is a standard procedure (Scott and Tribble 2006). For example, in order to find aboutgrams of an engineering research article, Warren (2010) first extracted its most frequent concgrams and then compared them to two reference corpora: a specialised corpus of engineering texts and a general reference corpus. Those concgrams which were more frequent in the original article than in the two comparison corpora were taken to be the text’s aboutgrams (Warren 2010). The procedure I am using in this study is very similar; therefore, some of the ‘overused’ patterns or even Non-matching concgrams are likely to be representing the specificity of the content of the text, i.e. its ‘aboutness’ or ‘keyness’, rather than idiosyncratic features of L2 use, personal style or usage habits of the author. I will call such patterns ‘content-related’.

There are other Matching but ‘overused’ concgrams in Kaisa’s and Maisa’s texts which look like candidates for the role of aboutgrams. For example, all of these concgrams appear just once in Kaisa’s C2 while being relatively frequent in her C1:

- APPLICATIONS/NLP (9/1)
- BACK/VOWELS (4/1)
- CASE/STUDY (7/1)
- FAMILY/TREE (5/1)

- FRONT/VOWELS (5/1)

- LINGUISTICS/THEORETICAL (10/1)

- OPEN/SOURCE (7/1)

- LANGUAGE/SPOKEN (13/1)

- CHANGES/SYSTEMATIC (8/1)

- LENGTH/VOWEL (4/1)

I will return to aboutgrams once again in Section 5.4.1, when analysing the reasons behind the Non-matching concgrams.

5.4. Non-matching patterns

The analysis of Matching but 'overused' patterns has conveniently laid ground for the discussion of Non-matching concgrams: here we also find individual preferences and content-related aboutgrams, even though preferences become even more specific and the keyness of aboutgrams increases.

As has been mentioned in the discussion of 'overused' patterns, the methodology of comparing one text to a corpus of other texts in fact answers the question how this text is different from other texts. The ways in which the text under examination will stand out will depend on the nature of the texts in the comparison corpus and how they are matched to the investigated text in terms of register, genre, language variety, authorship and other characteristics. Since in this study effort was made to find the kind of texts which matched C1 as closely as possible, to serve as data revealing priming rather than textual variation, there are not many features which distinguish between C1 and C2. However, inevitably, there are corpus-specific features which can have an impact on the Non-matching patterns identified. First, C1 is a corpus of drafts of one and the same text, thus some patterns as we have seen in the previous section show what this particular text is about. Second, C1 is a corpus of a Master's thesis, while C2 contains texts of other written academic genres. Therefore the comparison may reveal the kind of patterns which distinguish Master's thesis as a genre from other written academic genres. And finally, C1 corpus is a corpus of one author, making the comparison reflect this particular author's individual stylistic preferences and usage habits in addition to the patterns implicitly (or explicitly) learned from exposure.

Indeed, after all proper names are removed, the rest of the Non-matching concgrams fall into one of the following categories, strongly emerging from the data analysis: (1) patterns characteristic of the content, aboutgrams and wider context-induced patterns, (2) patterns characteristic of the genre and (3) patterns characteristic of the language use(r), preferences and approximations. While content-related and genre-specific patterns are natural features of any text, irrespective of its author and the L1 of the author, patterns characteristic of the language use(r) or individual preferences are the patterns which can reveal some regularity typical of L2 acquisition and use in addition to explicit, more conscious and considered lexical choices and language preferences. We will look at each of these categories in the corresponding sections that follow.

5.4.1. Content-related patterns

I will start with the patterns characteristic of the content. Aboutgrams, briefly introduced in Section 5.3.3, are patterns of this kind. Some of them tend to be terminological in nature and therefore present quite clear cases of aboutness. The concgrams Table 5.9 displays seem to be aboutgrams for Kaisa's and Maisa's C1s (for the full lists, see Appendix B): all of these concgrams have zero occurrences in the respective C2s.

Table 5.9 Aboutgrams from Maisa's and Kaisa's C1s which do not occur in their C2s

Maisa			Kaisa		
Concgram	n in C1	the pattern it represents	Concgram	n in C1	the pattern it represents
DEPENDENCY/RATIO	8	*dependency ratio*	DAUGHTER/LANGUAGES	8	*daughter languages*
ANNUAL/GROWTH	4	*annual growth*	ENVIRONMENTS/PHONETIC	5	*phonetic environments*
CHILDBEARING/TEENAGE	9	*teenage childbearing*	INVENTORIES/PHONEMIC	5	*phonemic inventories*
DOUBLING/TIME	8	*doubling time*	LANGUAGE/LIVING	7	*a living language*
HIGH/VARIANT	4	*high variant*	PROTO/WORDS	7	*proto-words*
LABOUR/MIGRATION	6	*labour migration*	ONCE/SPOKEN	6	*once-spoken language*
LOW/VARIANT	5	*low variant*	RULE/VARIABLES	10	*rule-variables*
MEDIUM/VARIANT	7	*medium variant*	GRAMMARS/TWO	6	*two-level grammars*
PLACE/RESIDENCE	5	*place of residence*	ETYMOLOGY/WORDS	4	*words with* X *etymology*

While, strictly speaking, an aboutgram is a co-occurrence of two words which is key to the text or collection of texts in question and therefore is characteristic of the content of this text in some way, some concgrams, which can also be called aboutgrams, represent wider context-induced patterns. For example, Maisa's research design requires her to describe different scenarios. As a result, she develops specific usage patterns which become more and more fixed. There are 16 concgrams which reflect this theme of her thesis. They are not aboutgrams in their purest form, because they are not meaningful themselves, they only point to the larger pattern which is key to the text. These 16 concgrams can be roughly divided into two types: those which categorise different scenarios and those which reflect the grammatical structure Maisa uses to consider and compare these scenarios, the second conditional.

The concgrams of the first type are: DECLINING/SCENARIOS (11/0), ART/NO (10/1), NO/SCENARIO (12/1), CONSTANT/HIV (32/0), CONSTANT/SCENARIO (15/0), CONSTANT/DECLINING (11/0), CONSTANT/SCENARIOS (10/0). The patterns they represent are: *constant HIV scenario, declining HIV scenario, constant HIV and declining HIV scenarios, No ART(-)scenario, No HIV scenario, No AIDS scenario.* There are also *hypothetical scenario* (n=3), *Full medication scenario (n=2)* and *Partial medication scenario* (n=1*)*, but these terms did not pass through the statistical thresholds set.

The concgrams representing the grammatical structure key to describing the different scenarios are: IF/WERE (3/3), ANTIRETROVIRAL/IF (5/0), ANTIRETROVIRAL/THERE (4/0), IF/TREATMENT (6/0), IF/THERE (12/9), IF/NO (14/4), NO/TREATMENT/6/2. It is clear that most of the concgrams refer to one and the same pattern. The concgram IF/NO is the most frequent one, thus, I chose it for the query presented in (5.33).

(5.33)

1 UN are provided as a default and can be used **if no** other data are available. In this projection,
2 200,000 more people would be alive without it. **If no treatment were** given during the whole
3 the population size would be around 2 700 000 **if there were no** AIDS, and slightly under 2 000
4 of AIDS on life expectancy at birth is dramatic. **If there were no** AIDS, the life expectancy in 2011
5 15–64 year-olds of total population is growing. **If there were no** HIV AIDS, the dependency ratio
6 with time, being about 70 years in 2020. **If there were no** HIV AIDS the doubling time would be
7 thus are not visible as separate in the graph. **If there were no** AIDS, the population would be
8 reducing fertility is also very possible here. **If there were no antiretroviral treatment**, **there**
9 antiretroviral treatment can be detected here. **If there were no** ART, the prevalence for 15–49
10 the amount of AIDS deaths, which would increase **if there were no treatment**. Consequently, the
11 time. The number of births would decline more **if there were no antiretroviral treatment**. The
12 is growing more rapidly than rural, as expected. **If there were no** AIDS, the growth would be a bit
13 is a "No ART" -scenario, showing the impacts **if there were no antiretroviral treatment**. In this
14 No ART -scenario is also shown in figure 14. **If there were no** ART, the population would naturally

(Maisa, C1, concordances for the concgram IF/NO)

Example (5.34) provides a quote from the text which explains what the scenarios mean and what they are used for:

(5.34)

> For the population projection, two scenarios were made. Both follow the HIV prevalence curve made in EPP until year 2015. In 2016–2020, the first scenario assumes that HIV prevalence is staying constant and the second one that it is declining. These scenarios are also compared to two hypothetical scenarios, the first one presenting a situation with no HIV/AIDS. In the second hypothetical scenario, there would be no antiretroviral treatment and the HIV/AIDS situation would be like in the "constant HIV" scenario.

(Maisa, C1)

So it is natural that the second conditional is used for describing certain scenarios as they are hypothetical. Although some concgrams appear in C2, the whole pattern is not represented there.

In C2 scenario(s) is/are also used but not that frequently, as can be seen from the concordance lines in (5.35).

(5.35)

1 not been affected by the HIV AIDS epidemic. This **scenario** is developed by removing estimates of AIDS
2 V in this age group. It is possible to avoid this **scenario** if prevention efforts are rapidly rolled out.
3 rtality levels and trends under the hypothetical **scenario** of no epidemic, then adds estimated AIDS-
4 del fits, nor can it estimate high and low future **scenario**s for the HIV epidemic based on the parameters
5 erent assumptions on outcome measures in various **scenario**s. For example, it is easy to vary levels of
6 NAIDS among others. A hypothetical "Without-AIDS **Scenario**" is created using RUP to model what would
7 e mortality results under both the "Without-AIDS **Scenario**" and the "With-AIDS Series" for Malawi. The

(Maisa, C2, concordances for the word *scenario(s)*)

Another example of a wider context-induced pattern comes from Kaisa's writing. The pattern which is frequent in her writing but does not occur in her C2 is *in the Alphabet* (5.36). More specifically, she is talking about *declaring, describing, defining* and *pairing up* different sounds and characters *in the alphabet*.

(5.36)

1	**Defining** the **archiphonemes in the alphabet** only, and not restricting the	kaisa_9_300410.txt
2	**describing** a set of **archiphonemes in the Alphabet** part of the grammar. If	kaisa_9_300410.txt
3	s /p/ and /k/, I can **declare in the Alphabet** of the Sumerian grammar **b:p** a	kaisa_9_300410.txt
4	**character pairs** have to be **declared in the alphabet**. In the traditional TwolC	kaisa_7_021209.txt
5	means that I **define in the Alphabet**-section of the grammar both **ä:a and a:a**	kaisa_8_130410.txt
6	you **declare** also **i:i** and **j:j in the alphabet**, HFST-TwolC concludes that i	kaisa_7_021209.txt
7	eans that if you **declare i:j in the alphabet**, unless you declare also i:i	kaisa_7_021209.txt
8	declaring an ambiguous mapping **in the Alphabet** causes generation	kaisa_8_130410.txt
9	Having to **declare the stops in the Alphabet** might seem a bit misleading,	kaisa_9_300410.txt
10	**sound changes** are covered with a different symbol **in the Alphabet**. In suc	kaisa_9_300410.txt
11	that needed to be **paired up in the Alphabet**. In such a manner one could	kaisa_9_300410.txt

(Kaisa, C1, concordances for *in the alphabet*)

The pattern *in the alphabet* does not occur in C2, but at least one example from it (5.37) shows that what Kaisa is doing is a normal procedure in her field, also in terms of word choice:

(5.37)

> We define an alphabet of special symbols that contains a unique symbol for each of the symbols in the original string alphabet.

(Kaisa, C2)

Admittedly it is not always possible to reliably distinguish between aboutgrams which are, so to speak, the preferences of text, and individual preferences of the language user. But it is probably unnecessary to draw a strict dividing line between aboutgrams and individual preferences, as this is not the intention here. The main point is to show that one of the reasons behind Non-matching patterns is the specificity of text which demands the use of specific terminology.

5.4.2. Genre-specific patterns

Even though both C1 and C2 texts are taken from academic discourse, there is an important genre difference between them which has to be taken into account and, as will be shown, has a bearing on patterns. C1 is a corpus of Master's thesis drafts, while C2 is a corpus of academic publications in the field: it consists of journal articles for the most part, but also articles from edited volumes.[64] As Hyland writes, "[b]oth PhD and master's dissertations are high-stakes genres for students… They carry the burden of assessment and determine future life choices, but with different expectations for particular forms of argument, cohesion and reader engagement. The problem for master's students is to demonstrate a suitable degree of intellectual autonomy while recognising readers' greater experience and knowledge in the field" (Hyland 2008: 47). In contrast, "[t]he primary social function of academic papers is to contribute to the goals of scientific inquiry" (Mauranen 1993: 19). In short, while a Master's thesis is a specimen of apprentice to experts writing, an academic article represents the writing of experts to experts. The following examples of Non-matching concgrams from

[64] Some exceptions: Hertta's C2 includes one PhD and one MA thesis, Maisa's corpus contains several reports prepared for or by e.g. the World Health Organisation or the Ministry of Health and Social Services in Namibia, Linda's C2 includes a press release and two magazine articles. All there rather different types of texts nevertheless constitute important sources of priming for the respective students.

Maisa's (5.38) and Kaisa's (5.39) data sets present some patterns which can be regarded as genre-specific:

(5.38)

- MY/PROJECTION (9/0)
- PREVIOUS/STUDIES (5/0)
- PREVIOUS/PROJECTIONS (7/0)
- I/WILL (11/0): *I will use/introduce/describe/summarize*
- I/USE (5/0)

(Maisa, C1)

(5.39)

- I /WILL (27 /2)
- FIRST /WILL (7 /2): *first I will, I will first* (both of the C2 occurrences are unrelated)
- SECTION /WILL (5 /1): *section will, In this section I will*
- CONCENTRATE /WILL (5 /0): *I will concentrate*

(Kaisa, C1)

Certain features are common to all of these examples and at the same time separate them from the patterns typical of published academic writing. The use of personal pronouns *I* and *my* is quite normal for theses but is often avoided in academic journal publications, especially if they are written by more than one author. Heavy emphasis on the earlier work done on the topic (*previous studies, previous projections*) is also something which is expected from a student: one needs to show what s/he has learned. Genre differences are likely to be reflected in many other lexico-grammatical patterns of the two corpora, but these are the ones which became visible through the comparison of concgrams.

At the same time, a lot of the genre-specific patterns are also metadiscoursal in nature, a territory where authors seem to be especially prone to developing their own preferences for wording since in metadiscoursal comments the voice of the author is most prominent, and, therefore, they reflect the personal authorial style of the writer. Also, as will be pointed out in Section 5.5.2, metadiscoursal comments tend to become very fixed due to constant repetition. Kaisa's predictive organising pattern emerging from the examples given above is an example of both her stylistic preference and its fixing, in addition to genre-specificity of some of its features.

Yet, in spite of the above said, it was decided to group such patterns under genre-specific patterns for several reasons: (1) C1 and C2 corpora are comprised of texts from different genres, (2) there are obvious genre-specific features in the patterns, and (3) it is likely that if Maisa and Kaisa were writing an article or a book rather than a Master's thesis, the (metadiscoursal) patterning they would be using would also be different.

5.4.3. Individual preferences

Content-related and genre-specific patterns are the kinds of patterns we expect to find in any text: it is common knowledge that content and genre of a text determine its lexical profile to a certain degree. In contrast, the patterns that will be focused on in this section seem to be individual preferences of a language user. Since writers in this study are L2 users, the patterning they display can reveal certain features and processes which pertain to L2 acquisition and use.

On the one hand, the patterns that will be discussed here seem to be non-matching to reference corpus because of a certain specificity of the lexical choice. On the other hand, they can be characterised by an unusual level of fixedness: they can be described as units of meaning which have more fixed components than are usually attributed to them.

Many of the patterns characterised as individual preferences are similar to the matching but 'overused' patterns. The cases of Maisa's preference for *antiretroviral treatment* instead of *antiretroviral therapy* and an anticipatory *it*-construction instead of a subject-to-subject raising construction just happened to have a corresponding pattern in her C2 and therefore fell into the Matching category. Since all the cases that will be discussed in this section are not ungrammatical in any way and are otherwise legitimate, it might be accidental that they do not have any matching patterns in C2. At the same time a fact of preference is undeniable, and this is what is interesting for us here.

We will look at just one example from Maisa's C1 because the already mentioned cases of *antiretroviral treatment* vs. *antiretroviral therapy* and an anticipatory *it*-construction vs. subject-to-subject raising construction illustrate exactly the same type of patterning: individual preferences. One of Maisa's favoured wordings which does not have a precisely corresponding pattern in her C2 is represented by the concgram CAN/DETECTED (9/0), the concordances for which are given in (5.40).

(5.40)

1 the time period where TFRs are available, it **can be detected** that the TFR has been declining at a
2 used where necessary. The **impacts** of HIV AIDS **can be detected** in population growth and population
3 for 2021. Also here the **impact** of HIV AIDS **can be detected**, although it remains unclear, how
4 have less children. Here the **effect** of education **can be detected**, but it is the proximate
5 data. Also the trends between 1992 and 2007 **can be detected**, as they are needed as basis for the
6 the **consequences** of antiretroviral treatment **can be detected** here. If there were no ART, the
7 ART. The demographic **impact** of the treatments **can be detected** here as in some of the following
8 the age structure of deaths, the **impact** of AIDS **can be detected**. In figure 18, deaths are
9 HIV prevalence among women, discussed earlier, **can be detected** in the projection results. In the

(Maisa, C1, concordances for the concgram CAN/DETECTED)

It can be noticed that the pattern is not only characterised by the fixed component *can be detected*, which is in itself noteworthy (the lemma DETECT is used in a grammatically different pattern just once: *is detected*),[65] but also by a semantic preference for different kinds of 'impacts' that are detected (*impact(s), effect, consequences).* It might also be important that it is the 'impact' *of HIV, AIDS, HIV prevalence, (antiretroviral) treatment(s)* or in one case (probably by an extension of the pattern) *of education* that is *detected.*

The above pattern does not have a matching equivalent in C2. It is not that the verb DETECT is not used at all in C2. In fact, it has 17 occurrences, but it does not have this fixed association with the passive voice and the modal verb *can* which happens to be Maisa's preferred wording. *Errors, infection, epidemic, trends* but also things like *age misreporting* are often *detected* in C2, thus forming a much looser group of co-occurring items than the ones Maisa has a preference for. A look at the noun collocates of the verb DETECT in the BNC suggests that it does not have one clear semantic preference which probably means it participates in more than one unit of meaning, yet one of its semantic preferences seems to be for 'medical phenomena' like *antibody/ies, dna, cancer(s), cells* (all from the 30 most significant collocates of the lemma). So the verb's high frequency in both Maisa's C1 and C2 is probably not accidental.[66] In the BNC, it also participates in a collocation with *error(s)* and *presence of,* most prominently. In all, DETECT in Maisa's usage seems to have many more syntagmatic associations which also have a set-like quality in comparison to the language use in her source texts. These syntagmatic associations constitute her individual preference.

[65] act of HIV on fertility is twofold. First, HIV **is detected** to reduce fertility, also when contraception use

[66] In the BNC the lemma DETECT occurs ca. 34 times per 1M words which makes 0.7 per 20k in comparison to Maisa's 10 per 20k and her C2's 4.25 per 20k.

Just like in Maisa's case, we have already noticed traces of developing preferences in Kaisa's examples too, for instance in the predictive organising pattern of her metadiscoursal comments, which are also becoming quite fixed and therefore will be discussed in more detail in Section 5.5.2. Here we will look at several other examples of her individual preferences.

Kaisa writes about relatedness between languages and chooses to use the words *affinity* or *affiliation* either on their own or in collocation with *linguistic* or *language*. LANGUAGE (LINGUISTIC)/AFFINITY (AFFILIATION)[67] could be considered aboutgrams, but it seems that expert writers in her field have a preference for *language relationship(s)* (5.41).

(5.41)

1 the program of an automated reconstruction of **language relationships** is completed. Keywords:
2 the program of an automated reconstruction of **language relationships** is completed. This is
3 chose to report the evaluation results for one **language** pair, since the **relationships** between the
4 for many African and native American **language**s, especially in the cases where the **relationship**
5 for measuring the degree of **relation** among **language**s. He used a core vocabulary of 115 basic
6 be useless for establishing **relationships** among **language**s. This may not hold true for young
7 in the cases where the **relationship** between **language**s has not been adequately proven. In
8 for a deeper analysis of **relationships** among **language**s. The point is that a tree is only an
9 of phylogenetic **relationships** between **language**s. However, this problem has received
10 able to obtain the **relationship** for any pair of **language**s. His conclusion is famous: La langue e

(Kaisa, C2, concordances for the concgram LANGUAGE/RELATION*)

In contrast *affiliation*[68] does not occur in C2. At the same time, it might be taken into account that in a further article Kaisa referred to but which was not included in the final version of C2 due to conversion difficulties, the word *affiliation* is used 4 times in the collocation *genetic affiliation* which does not show in the concordances. This additional article will be mentioned once more in what follows. As for *affinity,* it has just one occurrence in C2 (5.42).

(5.42)

or false friends depends on a certain **affinity** between the alphabets of the two languages

(Kaisa, C2)

It seems also that in Kaisa's own writing, *affinity* tends to co-occur with *language* (5.43) and *affiliation* with *linguistic* (5.44).

[67] In the BNC, no occurrences of any of the concgrams from the pattern LANGUAGE (LINGUISTIC)/AFFINITY (AFFILIATION).

[68] The concgram AFFILIATION/LINGUISTIC (4/0) prompted a search for patterns.

(5.43)

1	lso expectable, since the alleged **affinity** is a distant one. Since two-level	kaisa_9_300410.txt
2	provide evidence for the assumed **affinity**. Hence, the only way to prove gene	kaisa_1_170309.txt
3	which similar theories of **language affinity** can be tested. It worth emphasizi	kaisa_8_130410.txt
4	ng the actual theories of **language affinity**. Computational linguistics is no	kaisa_4_100809.txt
5	ht or wrong, or make the **language affinity** look more plausible, but to object	kaisa_8_130410.txt
6	ng the actual theories of **language affinity**. In the following I will provid	kaisa_10_300410.txt
7	not account for a true **linguistic affinity**. Case study – Sumerian:	kaisa_3_250609.txt
8	incide, the more reliably can the **affinity** be demonstrated via accurate and w	kaisa_1_170309.txt
9	related, the task of proving the **affinity** is tough since Sumerian is thousands	kaisa_1_170309.txt
10	ago, so the task of proving their **affinity** is challenging. The problem with r	kaisa_2_140509.txt

(Kaisa, C1, concordances for the word *affinity*)

(5.44)

1	rules in describing **language affiliation**, this section should rather b	kaisa_10_300410.txt
2	sk of finding true **linguistic affiliation** goes way beyond comparing	kaisa_1_170309.txt
3	econstruction: The **linguistic affiliation** of Sumerian. The first part o	kaisa_4_100809.txt
4	a novel theory of **linguistic affiliation** of Sumerian, a theory which	kaisa_4_100809.txt
5	hout considering the possible **affiliation** of the word. This is especial	kaisa_5_140909.txt
6	s on the theory of Sumerian's **affiliation** to Uralic, more precisely Fi	kaisa_5_140909.txt
7	nd new theories of **linguistic affiliation** are approached with grave res	kaisa_5_140909.txt

(Kaisa, C1, concordances for the word *affiliation*)

This fact may be pointing towards the splitting of a semantic preference type of association into two collocations, that is, a more abstract association becoming verbatim, or more fixed. In this sense it seems to be another instance of the process of fixing which will be taken up in Section 5.5.2.

Another Non-matching concgram which is related to these patterns is GENETIC/RELATIONSHIP (11/0) displayed in (5.45).

(5.45)

1 likely cognates. Borrowing by no means proves a **genetic relationship** so loans have to be eliminated from
2 Similarity Not all similarity is due to a **genetic relationship** between languages. [ref.]
3 Although word similarity alone does not prove a **genetic relationship**, and although the possibility of
4 feasible candidates is high enough to suggest a **genetic relationship**. The comparative method and
5 To what extent is it possible to evaluate a **genetic relationship** between two languages with the given
6 that especially in trying to define a distant **genetic relationship**, there are several other factors
7 is to demonstrate how a hypothesis on a distant **genetic relationship** can be formulated and tested, I
8 eems most likely that there is a more distant **genetic relationship** between Altaic and Sumerian as well.
9 possible cognates for determining a distant **genetic relationship**. Known history of the languages
10 means to prove the theory of the languages' **genetic relationship** right or wrong. Taken that the
11 relationship right or wrong. Taken that the **genetic relationship** of Sumerian and Finnish has not

(Kaisa, C1, concordances for the concgram GENETIC/RELATIONSHIP)

Again, it is a fixed contiguous collocation, sometimes also collocating with the adjective *distant* and the verbs PROVE or SUGGEST. Although, as we have seen, *relationship* is a frequent word in C2, it does not form a collocation with *genetic*. It must be noted, though, that the same article excluded from the corpus mentioned *genetic relationship* twice (5.46).

(5.46)

es but not by the rest may be evidence of close **genetic** relationship, continued contact, or typological
produce a more refined characterization of the **genetic** relationship among the languages. On one hand,

(Kaisa, a reference article outside her C2)

Taking into account that together with the phrase *genetic affiliation,* which appeared in the article 4 times, the combination *genetic affiliation/relationship* had in all 6 instances of occurrence in the article in question, it is possible to assume that this article served as a source of acquisition for the pattern in some way.[69]

In sum, we have three new units which are closely connected to the priming language of the field but are characteristic of Kaisa's usage in particular: *language affinity, linguistic affiliation* and *(distant) genetic relationship*.

Kaisa's concgram AID/NLP (5/0), in its turn, presents a developmental pattern which is very similar to the one observed in the case of CAN/DETECTED from Maisa's data set. Example (5.47) provides concordance lines for the concgram.

(5.47)

1 of historical linguistics has been used **to aid NLP applications** or vice versa. This section
2 cognate lists, automatically or manually collected, **aid NLP applications** like sentence alignment
3 of historical linguistics has been used **to aid NLP applications**. Most of the work done fal
4 A typical use for a generator as well is **to aid** other **NLP tools**, only in the opposite direct.
5 The goal of automatic cognate recognition is **to aid** other **NLP applications** like sentence alignment

(Kaisa, C1, concordances for the concgram AID/NLP)

It is interesting that in Kaisa's production the verb AID collocating with *NLP* occurs only in the infinitive (4 occurrences) and in the active voice (once, line 2), but not in any other word forms. As for C2, the pattern itself does not occur, but *NLP* is used 4 times, out of which once in a collocation with *applications* and twice with *tasks.* Further exploration of the pattern in C1 reveals that the concgram AID/NLP is part of a more general pattern, the pattern for the verb AID (see 5.48). The verb itself is almost exclusively (except for line 6) used in one and

[69] The BNC patterning: *genetic affiliation* - 0 hits, *genetic relationship* – 7 hits per 100M, compared to 6 hits of *genetic affiliation/relationship* per 60k words in C2.

the same form, as an infinitive, but with different objects. The pattern can be expressed in the following way: *to aid historical linguistics/language learning/NLP applications/NLP tools.*

(5.48)

1	computational historical linguistics does not seem **to aid historical linguistics**	kaisa_10_300410.txt
2	computational historical linguistics does not seem **to aid historical linguistics**	kaisa_4_100809.txt
3	use cognates of closely related languages **to aid in machine translation** or se	kaisa_10_300410.txt
4	the sake of preparing cognate lists **to aid language learning**. According	kaisa_summary_210109.txt
5	for the sake of preparing cognate lists **to aid language learning**. The authors	kaisa_10_300410.txt
6	cognate lists, automatically or manually collected**, aid NLP applications** like	kaisa_10_300410.txt
7	knowledge of historical linguistics has been used **to aid NLP applications**.	kaisa_4_100809.txt
8	knowledge of historical linguistics has been used **to aid NLP applications** or	kaisa_10_300410.txt
9	The goal of automatic cognate recognition is **to aid** other **NLP applications**	kaisa_4_100809.txt
10	A typical use for a generator as well is **to aid** other **NLP tools**, only in t	kaisa_6_300909.txt

(Kaisa, C1, concordances for the verb AID)

In contrast, in C2 the only occurrence of the lemma AID comes from a title of an article: "Bi-Textual Aids for Translators".

It seems that a frequently used item tends to develop and accumulate different kinds of associations extending its pattern. A search for *NLP applications* in (5.49) shows that it has obtained usage patterns of its own too. In addition to the verb AID, it also collocates with the verb REFINE. The context of occurrence can be expressed as: something *used to (help) aid/refine NLP applications*.

(5.49)

1	lists, automatically or manually collected, **aid NLP applications** like sentence	kaisa_10_300410.txt
2	ge of historical linguistics has been **used to aid NLP applications**. Most of the	kaisa_4_100809.txt
3	e of historical linguistics has been **used to aid NLP applications** or vice versa.	kaisa_10_300410.txt
4	automatic cognate recognition is **to aid** other **NLP applications** like sentence	kaisa_4_100809.txt
5	hand, using linguistic information **helps refine NLP applications**. In the	kaisa_10_300410.txt
6	edge of historical linguistics is used **to refine NLP applications**. The most	kaisa_10_300410.txt
7	r hand, using linguistic information **helps refine NLP applications**. The same	kaisa_4_100809.txt
8	s the computational implementations. Typically, **NLP applications** used in	kaisa_10_300410.txt

(Kaisa, C1, concordances for *NLP applications*)

What we can conclude on the basis of these examples is that individual usage can be characterised by units of meaning which have evolved from the priming language but have developed special features of their own. New units of meaning e.g. *language affinity, linguistic affiliation* and *(distant) genetic relationship* evolve from the loose usage *language relationship(s)/relationship(s) between/among languages*. Or units of meaning acquire new components, e.g. the lemma DETECT develops colligations with passive voice and modality and a semantic preference for different kinds of 'impacts' that can be detected, *NLP applications* develop a collocation with the verbs AID and REFINE.

Likewise, there are patterns whose high level of fixedness is the only trait which makes them different from reference corpus patterns. We will look at two examples, one from Maisa's writing and a similar one from Kaisa's C1.

Maisa's 'overly' fixed pattern has come up in four concgrams, two of which do not have any corresponding co-occurrences in her C2: NOT/MUCH (7/1), MUCH/OTHER (5/0), DIFFER/FROM (5/6), DIFFER/MUCH (5/0). To find all the occurrences of the pattern, I generate the concordance lines for the concgram DIFFER*/FROM, displayed in (5.50).

(5.50)

1 the behaviour of these two populations often **differ**s **from** one another. 2 The research setting The
2 Both scenarios of this projection, which **do not differ much from each other**, fit quite in the middle o
3 HIV" scenario. The scenarios **do not** finally **differ much from each other**. The hypothetical scenario
4 HIV scenario. The two scenarios **do not differ** here **much from each other**, and thus are not
5 explain the difference. Also, UN's assumptions **differ** quite **much from** my assumptions, for example in
6 Constant HIV and Declining HIV scenarios **do not differ** quite **much from each other**. This is of course

(Maisa, C1, concordances for the concgram DIFFER*/FROM)

As can be seen, these concordance lines capture all the instances of all the concgrams in question. The same operation is then performed with C2 (5.51):

(5.51)

1 or to work-based clinics. These people may **differ from** the general population in significant
2 refuse to participate in sentinel surveillance **differ from** those who agree to participate. This bias
3 surveillance is generally carried out may **differ from** those who attend private clinics or who do
4 Fertility—Women who become pregnant may **differ from** women who do not become pregnant in ways
5 surveillance issues in generalized epidemics **differ** somewhat **from** those of low-level and concentrated
6 IDB are prepared using data and procedures that **differ** slightly **from** those used for all other

(Maisa, C2, concordances for the concgram DIFFER*/FROM)

The general pattern in both corpora is the same: it is the normal pattern in which the verb DIFFER participates, that is, X DIFFER *from* Y. Yet some specific features are also noticeable. In C2 (5.51), the verb is used in the positive form sometimes hedged by the modal verb *may.* Maisa uses the verb in the positive form too, but 4 out of 6 occurrences (lines 2, 3, 4, 6 in Example [5.50]) feature the pattern *do not (finally) differ (quite) much from each other*. The pattern comprised of 7 words is almost totally fixed: it consists of fixed word forms instead of lemmas and does not include any intervening words or optional associations of a more abstract nature. On closer examination, it turns out that it is always used to talk about scenarios. So in a way it also belongs to the ‘Scenario-pattern’ analysed above, in Section 5.4.1. The question which immediately springs to mind is whether all the instances of the pattern occur close to each other in the text. Indeed three instances of the pattern occur in the same draft but one is in a different draft. Thus it seems recency of use might be an important factor having a considerable impact on the level of fixedness this pattern exhibits.

Even though the pattern is not ungrammatical and represents a valid usage option for the verb DIFFER, the chances that somebody else uses it in such a fixed form are low. A query in the BNC of “_VM0 ** not ** differ” and “_VD* ** not ** differ” (where VM0 stands for any modal auxiliary and VD* for any form of the verb DO, while asterisks take into account possible intervening words) brings nine occurrences of the same pattern (i.e. co-occurring with *much*).[70] These occurrences, seven of which are provided in (5.52), prove that the pattern is valid and used, it has only become preferred and fixed in Maisa’s writing.

(5.52)

1 of statement. A statement, I'd say, that **might not differ much from** the one made 56 years ago by his
2 from equation (3.20), in a large sample**, should not differ very much from** the estimates from equation
3… until the mid-Triassic, although they probably **did not differ much from** their late Permian ancestors.
4… (see Appendix I, Table 12c) women and men **did not differ much** in their attitudes to different types of
5…in the childhood supplement. Age specific death rates **do not differ much** between the non-manual social
6… British school and college mathematics classrooms **do not differ much** from those of a hundred years ago.
7… Lashner et al reported a 3.2% incidence. Crohn's colitis **does not differ very much from** ulcerative colitis in

(BNC)

There is a similar example of an ‘overly’ fixed pattern in Kaisa’s production too (5.53): 8 concgrams, AN/FRONT (5/0), AN/EPENTHETIC (9/0), E/FRONT (6/0), E/EPENTHETIC (12/0), FRONT/I (5/0), FRONT/Y (5/0), EPENTHETIC/I (5/0) and EPENTHETIC/FRONT (5/0), all of which are Non-matching, represent one and the same pattern: *Y * (realised as) i * in front of an epenthetic e:*

(5.53)

1	plural nouns: y:i = **Y -> i in front of an epenthetic e** 0:e = An epenthetic	kaisa_6_300909.txt
2	this is expressed as" **Y realised as i in front of an epenthetic e**" Rule 2, on the	kaisa_6_300909.txt
3	2 separate rules: "**Y realised as i** only **in front of an epenthetic e**" Y:i => _ 0:e	kaisa_6_300909.txt
4	thetic e" Y:i => _ 0:e ; "Y realised as i always in front of an epenthetic e" Y:i	kaisa_6_300909.txt
5	but requires **Y** to always be **realized as i in front of an epenthetic e**. Such	kaisa_6_300909.txt

(Kaisa, C1, Y * (realised as) i * in front of an epenthetic e)

[70] It is interesting to add that both patterns "_VD* ** not ** differ" and "_VM0 ** not ** differ" co-occur with the adverb *significantly* more often that with *much,* 58 times out of 177 occurrences of the pattern in the first case and 3 out of 12 occurrences in the second.

It must be mentioned that again all the occurrences come from the same draft where they are also located very close to each other.

There is one more type of an individual preference: when C1 pattern does not match C2 patterning even though C2 actually contains a direct equivalent. The reason for this may lie in the changes an L2 user has introduced into the unit. These changes can be interpreted as departures from the standard form rather than individual preferences developed through repeated use. These departures from the form a user was exposed to can be explained by the frequency effects on the one hand and by the superiority of the memory for meaning over memory for surface structure on the other. That is, when the frequency of exposure is not high enough, the exact form of a unit is not remembered and is therefore approximated, but with the overall meaning being retained. In other words C1 pattern may be an approximation of a C2 pattern. The term approximation is adopted from Mauranen (e.g. 2005, 2009) and was discussed in more detail in Chapter 3 (Sections 3.5 and 3.6). I will return to the process of approximation in Section 5.5.1, but we will also look at one example here.

The example which follows in (5.54) can be interpreted as a lexical approximation: it is a Non-matching concgram from Maisa's list, representing a collocation *DHS stud(y/ies):*

(5.54)

1 means the census data [ref.], **DHS studies** [ref.]

2 level of fertility. According to the previous **DHS studies**, TFR has declined from 5.4 in 1992 to 3.6

3 Figure 1. Total Fertility Rate in Namibia in **DHS studies** of 1992, 2000, and 2006–07. Rates for

4 been collected in 2009. Unfortunately, this **DHS study** does not provide HIV testing data, like

5 are discussed in the Final Reports of the **DHS studies**. In the 2006-07 survey, the sample was a

6 does not provide HIV testing data, like many **DHS studies** in other developing countries. More

7 HIV epidemic is advancing. In addition, a new **DHS study** of Namibia is being prepared to be

8 from the 1991 and 2001 censuses, and the three **DHS studies**. Sex ratio at birth is also needed, and

9 subsequent ones in 2000 and 2006 07 (Measure **DHS**, 2009). Several **studies** on fertility have been

(Maisa, C1, concordances for the concgram DHS/STUD*)

When it is compared to expert usage (see Example [5.55]), it turns out that the standard form is *DHS survey(s),* not least because the abbreviation DHS actually stands for Demographic and Health Survey(s).

(5.55)

1 on breastfeeding practices is found in **DHS surveys** fielded in most of the countries as well

2 include those undertaken as part of ORC Macro's **DHS surveys**, the World Fertility Surveys (WFS), the

3 DHS in a number of countries, the international **DHS survey** programme intends to include AIDS modules

…

41 such as the Demographic and Health **Surveys** (**DHS**) contain data that can be used to track changes

42 primarily from demographic and health **surveys** (**DHS**). Additional data included in this report were

43 of data produced by the **DHS surveys,** and some **DHS** reports mention the role of premarital
44 are gathered primarily by **surveys** such as the **DHS**, the family health and contraceptive prevalence

(Maisa, C2, concordances for the concgram DHS/SURVEY*)

In all, C2 has 44 occurrences of *DHS survey(s)*, 1 occurrence of *DHS reports* and no occurrences of *DHS stud(y/ies).* Neither was I able to find any occurrences of *DHS stud(y/ies)* in Google or Google Books, let alone the BNC which is not specialised enough for such searches. Thus, it seems reasonable to argue that the unit *DHS survey(s)* was approximated semantically in that the word *survey* was substituted by the word *study* as they belong to the same semantic set. Another reason for this substitution is that *stud(y/ies)* is a more common item than *survey(s):* it is three times as frequent in the BNC (32386 vs. 9690 instances).

At the same time it is not impossible that Maisa had encountered *DHS stud(y/ies)* and therefore was primed for such use rather than approximated it herself. In this sense this example is very similar to the case of *antiretroviral treatment* vs. *antiretroviral therapy.* The standard form, *antiretroviral therapy,* gave way to *antiretroviral treatment* in Maisa's usage in spite of the predominance of the standard form in C2. Yet, two occurrences of *antiretroviral treatment* were documented in C2, therefore it was not entirely correct to talk about approximation in this case as Maisa could have been primed to use *antiretroviral treatment.* But approximation could be at issue here too. Probably approximation can operate at the cognitive level, when a language user approximates the form of a unit of meaning due to fuzzy memory, and at the level of language community, when a unit of meaning becomes approximated due to repeated cognitively-based approximations by individual members of the community and their spread through priming.

All the patterns discussed in this section have one thing in common: a very high level of fixedness. Whether we are talking about new units of meaning which evolve in individual usage or units of meaning which acquire new components or the new status for already existing components, they become more and more fixed, presumably through repeated use contextually required time after time.

An interesting question is whether this is a gradual process and, if it is, how gradual it can be. Theoretically, it is difficult to imagine that something can become very fixed from just one occurrence. At the same time in the present data there is no evidence of any initial variation before some kind of equilibrium is reached. The data suggests that already on the second occurrence a pattern is reproduced in a quite fixed form. Yet, again variation is not something which easily lends itself to observation through the methods of n-gramming and concgramming.

In any case, while approximation as a process seems to be supported by the data, fixing seems to be a phenomenon strongly emerging from the data analysis as another process behind the mechanism of the idiom principle. I will now discuss these two processes, approximation and fixing, separately in the section that follows.

5.5.Two processes behind the mechanism of the idiom principle

In Chapter 3, it was suggested that the production of a unit of meaning on the idiom principle can be influenced by a cognitively natural process of approximation. Through approximation verbatim associations, i.e. collocations, can become semantically or structurally fuzzier. The phenomenon of approximation has been noticed in English as a lingua franca use at the macrosocial level of discourse communities. In other words, while it is intuitively plausible that units of meaning are approximated at the cognitive level, so far approximation has only been observed across speakers, as a product rather than as a process. One lexical item, for example *so to speak* appears in an approximate form, *so to say,* in the language use of several different speakers in unrelated communicative events (Carey 2013) suggesting that it is not an idiosyncratic error but a systematic pattern. It is reasonable to assume that each of these language speakers has approximated the standard form *so to speak* due to, for example, memory constraints described in Section 3.6. Yet, we have not seen how it happens: how in spite of the priming, the item starts to occur in a different form but with the same meaning and what happens to it after repeated use. In the previous section, we have already looked at one example of a Non-matching pattern which could be explained by approximation. In this section I will present more evidence of approximation in the present data and try to track it down as a process.

Approximation is a process which was predicted from the start on the basis of previous studies and theoretical considerations presented in Chapters 2 and 3. At the same time, the empirical analysis strongly suggests the existence of a second process, *fixing*. Fixing seems to be a reverse process to approximation. In simple terms, while through approximation a pattern loosens up and moves from the category of collocation to the category of semantic preference or colligation, through fixing, in reverse, it becomes more fixed, that is, a particular variant from a semantic set or a grammatical class starts to be preferred and becomes a collocation. I will summarise the examples of fixing we have seen so far and discuss some more cases in more detail in Section 5.5.2.

5.5.1. Approximation

Approximation was suggested as a process through which variability is introduced into a unit of meaning in Sections 3.5 and 3.6. In this section, I will inspect the data for support or refutation of this suggestion. In Section 5.4.3, we have already seen one example of approximation, since approximation seems to be one of the reasons why C1 patterns do not always match C2 patterning. Yet, approximation is a process which is difficult to identify because it presupposes variation.

Previous studies (e.g. Mauranen 2005, 2009, 2011) have shown that approximations can be divided into lexical and structural substitutions. It was argued in Section 3.5 that lexical substitutions resemble a semantically approximated collocation, i.e. semantic preference, and structural substitutions resemble a grammatically approximated collocation, i.e. a colligation. Here we will look at one example of lexical and one example of structural approximation which arise from the comparison of C1 concgrams with C2 co-occurrences. In addition to this, I will show that approximation can also be detected using a different methodology.

An example of a lexical or semantic approximation is presented by the concgram MEDICATION/USE, which has five instances in Maisa's C1 (5.56), but no instances in C2.

(5.56)

1 Two other scenarios are about the extent of **medication use**. In the "Partial **medication**" scenario,
2 use. In the "Partial **medication**" scenario, **medication use** begins when an HIV-positive person becomes
3 the extent of **medication use**. In the "Partial **medication**" scenario, **medication use** begins when an
4 all people who become symptomatic, begin to **use medication** in 2006 (it is assumed that 30% of the peopl
5 government programs and the wider **use** of HIV **medication**. The first of these scenarios, "Behavioural

(Maisa, C1, concordances for the concgram MEDICATION/USE)

The word *medication(s)* is used in C2 but in different contexts, therefore it was decided to search for its close synonym *medicine(s)* as well. It is only used four times out of which twice in an unrelated context, as an explanation for an acronym (5.57):

(5.57)

1 ZT 3TC Zidovudine and lamivudine (anti retroviral **medicines**) DHS Demographic and Health Survey EPP
2 g and Evaluation NVP Nevirapine (anti retroviral **medicines**) PMTCT Prevention of mother to child
3 bia. These services ensure that women **receive** the **medicines** required to reduce the risk of transmitting
4 treatment because of the stigma, they stop **taking medicines** because of the side effects, or they default

(Maisa, C2, concordances for the word *medicine(s))*

But the two times it is used in the context relevant for the comparison, different verbs collocate with it: TAKE and RECEIVE. A search for the most significant verb collocates of *medicine** and *medication** in the BNC shows that indeed the most common verb that is used for this purpose is TAKE, the verb RECEIVE is also a significant collocate of *medication** though not of *medicine**. In contrast the verb USE, although appearing in the lists of significant verb collocates for both of the items, on closer examination functions in a different pattern: either when something is used in medicine/medication or when medicine/medication is used in general, e.g., for a certain purpose, but not by someone. Some examples of this pattern are provided in (5.58).

(5.58)

1 The same constitutional **medicine** can be **used for** both palliative and curative purposes simultaneously

2 that several bottles of **medicine** will be **used** in the process of cure

3 ancient China, herbal **medicine** was **used** in conjunction with acupuncture

4 Can homoeopathic **medicines** be **used** to treat all kinds of conditions?

5 personal experience that high technology **medicine** is **used** inappropriately

6 this was often the very same **medication** that was **used** in the subsequent overdoses

7 Wouldn't the **medications used** affect the beneficial bacteria?

(BNC, examples of *medicine*/medication** + the verb USE)

If we go back to Maisa's usage (5.57), we will see that actually in 4 out of 5 instances she employs the verb USE in exactly this general sense. However, in line 4, she repeats the pattern when she starts talking about people too. This can be viewed as a lexical approximation of a more specific patterning: a collocation of *medication(s)/medicine(s)* with the verbs TAKE/RECEIVE when the agents are people is approximated to a generalised collocation of *medication(s)/medicine(s)* with the verb USE.

Let us now move on to the structural or grammatical approximation. Maisa's non-matching concgram YEAR-OLDS/INCIDENCE (3/0) does not make sense as such but when concordance lines are queried (5.59), it turns out that it does indeed represent a pattern, though *incidence* in this pattern is not a compulsory component, it is more of a member of a semantic set with which the unit co-occurs:

(5.59)

1 reversed. Figure 7. **HIV incidence of 15–49 year-olds**,% When looking at incidence as number of people

2 the future. Figure 4. **HIV incidence of 15–49 year-olds**,% HIV prevalence and incidence as number of

3 thousands Figure 6. **HIV incidence of 15–49 year-olds**, thousands 8.2 Population projection Next, the

(Maisa, C1, concordances for the concgram YEAR-OLDS/INCIDENCE)

The concordance lines generated by a more general query for "*year-olds*" in both C2 and C1 are displayed in (5.60) and (5.61), respectively:

(5.60)

1 ne, a dramatic rise in death rates **among** 15 to 45 **year-olds** can provide an indication of excess
2 the contraceptive prevalence **among** the 15- to 24-**year-olds** (correlation coefficient was 0·61 with ever
3 2004 to 27.1% in 2005. HIV prevalence **among** 15–24-**year-olds** in Lesotho appears to be stabilizing. A
4 o ensure large enough sample sizes for the 15–24 **year-olds**. This will allow improved tracking of

(Maisa, C2, all instances of *year-olds)*

(5.61)

1 oncerning age structure. The age groups of 15–59 **year-olds** are becoming larger than before, and the
2 DS deaths are occurring in the age group of 30–39-**year-olds**. Changes in age groups 20–29 and 40–49 are
3 most deaths will occur in the age group of 30–39- **year-olds**. "Estimates and Projections of the Impact of
4 n the future. Figure 4. HIV incidence of 15–49 **year-olds**,% HIV prevalence and incidence as number of
5 59. With the Constant HIV scenario, adult (15–49 **year-olds**) HIV prevalence would be 13% in 2020,
6 The projected adult prevalence rates for 15–49 **year-olds** in 2020 are 13% in Constant HIV scenario
7 seen that the prevalence is lower than for 15–49 **year-olds** in all projection years. This is well
8 t projection years. Figure 8. Number of 15–49 **year-olds** infected with HIV 2.2 Population projection
9 0 (2021), an increase of 43%. The group of 40–49-**year-olds** is assumed to grow from 140 000 (2001) to
10 ation is decreasing while the proportion of 15–64 **year-olds** is increasing, as already shown in figure
11 he projection period. Figure 6. Number of 15–49 **year-olds** living with HIV, Namibia 1970–2015. EPP also
12 heir relative size, while the proportion of 15–64 **year-olds** of total population is growing. If there
13 increased in size, especially the groups of 15–59 **year-olds**. The impact of declining fertility rates is
14 9, 0.53, and 0.47 (from 20–24 **year-olds** to 45–49 **year-olds**). The ratio of the youngest age group can be
15 very slowly. Figure 5. HIV prevalence of 15–49 **year-olds**, thousands Figure 6. HIV incidence of 15–49
16 olds, thousands Figure 6. HIV incidence of 15–49 **year-olds**, thousands 8.2 Population projection Next,
17 76, 0.71, 0.65, 0.59, 0.53, and 0.47 (from 20–24 **year-olds** to 45–49 **year-olds**). The ratio of the
18 een reversed. Figure 7. HIV incidence of 15–49 **year-olds**,% When looking at incidence as number of
19 If there were no ART, the prevalence for 15–49 **year-olds** would fall below 8% in 2020. When looking

(Maisa, C1, all instances of *year-olds)*

Not all the instances of *year-olds* in (5.61) concern incidence/prevalence or some other rate in an age group and are in that similar to the initial concordance lines generated by the concgram YEAR-OLDS/INCIDENCE. Example (5.62) therefore shows the selected lines relevant for the comparison.

(5.62)

4 n the future. Figure 4. **HIV incidence of 15–49 year-olds**,% HIV prevalence and incidence as number of
6 The projected adult **prevalence rates for 15–49 year-olds** in 2020 are 13% in Constant HIV scenario
7 seen that the **prevalence** is lower than **for 15–49 year-olds** in all projection years. This is well
15 very slowly. Figure 5. **HIV prevalence of 15–49 year-olds**, thousands Figure 6. HIV incidence of 15–49
16 olds, thousands Figure 6. **HIV incidence of 15–49 year-olds**, thousands 8.2 Population projection Next,
18 een reversed. Figure 7. **HIV incidence of 15–49 year-olds**,% When looking at incidence as number of
19 If there were no ART, the **prevalence for 15–49 year-olds** would fall below 8% in 2020. When looking

(Maisa, C1, selected instances of *year-olds)*

When Examples (5.60) and (5.62) are compared, it becomes clear that while expert writers use the preposition *among* to indicate prevalence/incidence or some other rate for a certain age group, Maisa uses *for* or *of*. This substitution can be interpreted as a grammatical or structural approximation: an approximation inside a certain grammatical class, in this case in the class of prepositions.

As mentioned above, it is not easy to systematically identify cases of approximation. The reason for this is simple: approximation presupposes some kind of change, and it is difficult to retrieve automatically two units which would mean the same but would have different formal representation. In this thesis it was hypothesised that if the students' written production contained cases of approximation, at least some of them would surface in the non-matching concgrams. This is what indeed happened, even though the numbers of identified cases were small. However, it is highly likely that these are not the only cases of approximation the data contains, and other cases were simply not captured by the method employed: in fact, the most frequent or most significant co-occurrences may not be the best place to search for approximation. As argued in Sections 5.3.3 and 5.4.1, many of them are likely to be key for the texts and therefore too salient for the authors to be approximated.

So, to probe the ground just a bit further, I tried running one student's C1 through a different phraseological programme, a skipgram instead of a concgram (cf. Römer 2011). William Fletcher's kfNgram (Fletcher 2002-2012) is able to automatically generate lists of 'phrase-frames' where one word would be variable, which gives a different view of phraseological patterning. By looking through a list of 4-word phrase frames generated from Kaisa's C1 using kfNgram and comparing them to 4-word phrase frames generated from her C2, I found some examples of matches and mismatches, presented in Table 5.10.

Table 5.10 Examples of 4-word phrase-frames from Kaisa's C1 which include a case of approximation

Phrase-frame	Number of occurrences	Corpus
*it is * to*	17	C1
*it is * to*	19	C2
it is hard to	4	C1
it is difficult to	5	C2
it is important to	4	C1
it is important to	3	C2
it is necessary to	4	C2
it is possible to	9	C1
it is possible to	7	C2

As it transpires from the table, it seems that while writers in C2 use the frame *it is difficult to*, Kaisa uses *it is hard to* which can be viewed as an example of a semantic approximation.

All the cases of approximation discussed in this section occur in recurring patterns which had direct equivalents in the priming language but did not reproduce them exactly. This suggests that it is possible to acquire and use a pattern in an approximated form but on the idiom principle. Also, lexical and structural substitutions are indeed remarkably similar to the categories of semantic preference and colligation. Thus, it seems that the process of approximation can work inside a unit of meaning. Together, these two observations suggest that approximation is a process which lies behind the mechanism of the idiom principle.

5.5.2. Fixing

As mentioned at the outset of this section, while approximation is a process which was predicted from the start, the process of fixing strongly emerges from the data analysis. We already had a chance to observe this process in many examples discussed in the previous sections: in the example of *assumptions about* developing a new field-specific component of semantic preference for demographic rates; *rapidly* starting to collocate with *more,* colligate with the Progressive aspect and develop a semantic preference for verbs with a sense of changing up or down the scale; *antiretroviral treatment* becoming a preferred collocation in spite of an alternative *antiretroviral therapy*; the verb ASSUME getting fixed in the pattern: *it is/can be assumed that* which serves as a convenient opening of a sentence; the verb DETECT occurring almost exclusively in the pattern *can be detected* + a semantic preference for 'impacts'; LANGUAGE (LINGUISTIC)/AFFINITY (AFFILIATION) splitting into two specific collocations *language affinity* and *linguistic affiliation*; fixed contiguous collocations, such as: *distant genetic relationship* and *closely related languages;* the verb AID occurring only in the infinitive; *NLP applications* developing a collocation with the verbs AID and REFINE; the 'overly' fixed patterns *do not (finally) differ (quite) much from each other* and *Y * (realised as) i * in front of an epenthetic e.* These examples suggest that fixing can operate in three different ways: through fixing (1) the components of a unit of meaning can become more strongly associated with it, (2) they can change their category from semantic preference and colligation to collocation, i.e. become verbatim instead of abstracted, (3) a unit of meaning can acquire new components. In this section, we will look at some more examples of fixing and see whether we can observe these three routes through which a unit of meaning becomes more fixed.

The first example involves the following concgrams which attract attention due to their apparent interconnectedness: COGNATE/RECOGNITION (12/2), AUTOMATIC/COGNATE (9/1), COGNATE/IDENTIFICATION (5/26), AUTOMATIC/RECOGNITION (7/0). There are two observations which can already be made without a further look at the concordance lines. First, it seems that Kaisa and the expert writers have opposite preferences: Kaisa mostly uses *cognate recognition* and expert writers mostly *cognate identification*, though both parties also use the second variant some of the time. Second, there is one concgram which is non-matching, AUTOMATIC/RECOGNITION (7/0), indicating that probably Kaisa's preferred expression, *cognate recognition*, has acquired an additional element to its patterning. Kaisa's C1 concordance lines for the concgram COGNATE/IDENTIFICATION in (5.63) and for the concgram COGNATE/RECOGNITION in (5.64) indeed show that the extended pattern is *automatic cognate recognition* which has become markedly fixed:

(5.63)

1 reconstruction. Automatic cognate recognition **Cognate identification** has received more attention

2 linguistics fall under two categories: **cognate identification** and language in

3 focusing on the two main areas of research: **cognate identification** and establishing

4 is to show that the algorithm can be used for **cognate identification** and that it can be

~~5... **identification** and reconstruction. Automatic **cognate** recognition **Cognate identification** has~~

(Kaisa, C1, concordances for the concgram COGNATE/IDENTIFICATION)

(5.64)

1 identification and reconstruction. **Automatic cognate recognition Cognate** identification has

2 Especially the early approaches to **cognate recognition** used only orthographic

3 historical linguists profit from **automatizing cognate recognition**. Second, **cognate** lists,

4 metrics. Nowadays, the general tendency in **cognate recognition** is to lean towards empirical,

5 results from automated methods**. Automatic cognate recognition** can be roughly split into two

6 section concise, since even though **automatic cognate recognition** is possibly the biggest field

7 friends (faux amis). The history of **automatic cognate recognition** goes back to the early 90's.

8 early 90's. One of the earliest approaches to **cognate recognition** is reported in [ref.]. In th

9 measures. The few examples of **automatic cognate recognition** above provide a very brief

10 given word pair list. Methods used in **automatic cognate recognition** are not directly applicable to

11 are somewhat related. The goal of **automatic cognate recognition** is to aid other NLP

~~12 reconstruction. Automatic **cognate recognition Cognate** identification has received more attention~~

(Kaisa, C1, concordances for the concgram COGNATE/RECOGNITION)

A search for AUTOMATIC/COGNATE, presented in (5.65), reveals that there is also a third alternative to the expression: *cognate detection*, but that in any case *automatic cognate* collocates with *recognition* 7 times out of its 9 occurrences:

(5.65)

```
1         Review Computational Historical Linguistics Automatic Cognate Detection Data  In this section,
2        of both identification and reconstruction.   Automatic cognate recognition  Cognate has
3      report better results from automated methods.   Automatic cognate recognition can be roughly split
4          false friends (faux amis).  The history of automatic cognate recognition goes back to the
5              measures.        The few examples of automatic cognate recognition above provide a very
6      I leave this section concise, since even though automatic cognate recognition is  possibly the
7        of the work done falls under the category of automatic cognate detection. Usually the  aim of
8      from any given word pair list.  Methods used in automatic cognate recognition are not directly
9      but the tasks are somewhat related. The goal of automatic cognate recognition is to aid other NLP
```

(Kaisa, C1, concordances for the concgram AUTOMATIC/COGNATE)

So, it can be seen that although there are three possible alternatives for the phrase: *cognate recognition, cognate detection* and *cognate identification*, Kaisa exhibits an overwhelming preference for just one of them which as a result develops into a fixed pattern *automatic cognate recognition.* This can be taken as evidence of semantic preference for a set of semantically related words (*recognition, detection, identification)* developing into a collocation with a specific word (*recognition).*[71] Also, the pattern extends to include a collocation with *automatic*.

Another candidate example for fixing is brought to light by the concgrams CHANGES/POSTULATED (5/0) and POSTULATED/SOUND (8/0). Although they are non-matching, there is evidence that it is not an idiosyncratic use. First, the verb POSTULATE actually occurs in C2 once (5.66) and in a quite similar context.

(5.66)

```
ere; the algorithm would  then be more willing to postulate prefixes and suffixes than infixes.   4. The Full
```

(Kaisa, C2, concordances for the verb POSTULATE)

Second, there is an almost identical example of usage (5.67) from the article mentioned above, which was referred to in Kaisa's thesis but not included in C2:

(5.67)

The program is a research tool designed to aid the linguist in evaluating specific hypotheses, by calculating the consequences of a set of **postulated sound changes** (proposed by the linguist) on complete lexicons of several languages.

(Kaisa, reference article)

[71] This is admittedly difficult to claim on the basis of the concordance data only: numerically *recognition* is a preferred alternative but this may not be a decisive feature. Yet, in a word association task, the data which will be discussed in the next chapter, when prompted with the word *automatic* Kaisa responded with a collocational response: *recognition*, i.e. not *detection, identification*. So the preference may be indeed becoming verbatim.

Let us now look at Kaisa's usage pattern for the verb POSTULATE in (5.68).

(5.68)

1 Finnish vowel system is broader, I **postulate** the **vowels** of Finnish to the proto- kaisa_8_130410.txt
2 laryngeal fricative /h/. I **postulate** all of the Sumerian **fricatives** to the proto kaisa_9_300410.txt
3 have **voiced stops** either, so I do not **postulate** any to the proto-language. For the kaisa_9_300410.txt
4 one single **proto-sound** can be **postulated** instead of several sounds. Such exam kaisa_2_140509.txt
5 level rules, and also **sound changes postulated** in historical linguistics, depend on kaisa_8_130410.txt
6 found in each cognate pair. I have **postulated** the **proto-sounds** in Table 3 so that kaisa_3_250609.txt
7 Second, the **sound changes** he **postulated** in each cognate pair needed to be kaisa_6_300909.txt
8 regularity of the **sound changes** he **postulated**, he provided me with a list of 1671 kaisa_3_250609.txt
9 form. The **proto-sound** is **postulated** based on the phonetic properties of its refle kaisa_2_140509.txt
10 case study: modeling a set of **postulated** Sumerian and Finnish **cognates** with two- kaisa_10_300410.txt
11 consistent pattern. If the **proto-sound postulated** for (1) would have been for kaisa_3_250609.txt
12 implementation. The **postulated sound changes** are tested by means of two- kaisa_6_300909.txt
13 test the regularity of the **postulated sound changes**. When a computational kaisa_8_130410.txt
14 modeling the **sound changes** he **postulates**, not on etymological evaluation kaisa_8_130410.txt

(Kaisa, C1, concordances for the verb POSTULATE)

It is easy to notice that Kaisa is using the verb POSTULATE almost exclusively (except for line 10) with the words from the semantic set of 'sounds', with *sound changes* being the most common collocate. The patterning of an extended unit of meaning is emerging here with the verb POSTULATE obtaining a collocation and a semantic preference. Again, it can be said the components of a unit of meaning around the verb POSTULATE become more fixed.

It now seems important to give some examples of fixing in metatextual items since they seem to be especially susceptible to this process, as already mentioned in Section 5.4.2. I will take two concgrams from Kaisa's writing which have very similar communicative functions: CONCENTRATE/WILL (5/0), displayed in (5.69) and GO/THROUGH (6/0), displayed in (5.70).

(5.69)

1 alternations, and **in the following I will concentrate** on **describing** the first practical
2 and historical linguistics, but **first**, **I will concentrate** on **demonstrating** the need to formalize
3 spoken, everyday language. **This section will concentrate** on **describing** the use of computational
4 and historical linguistics but **first I will concentrate** on **demonstrating** the need to
5 Proto-Finno-Ugric. **This example case will** only **concentrate** on **reconstructing** the initial sounds.

(Kaisa, C1, concordances for the concgram CONCENTRATE/WILL)

(5.70)

1 Sets and Definitions. **In the following I will go through** an example grammar from [ref.]
2 with Two-level Rules **In this section, I will go through** the steps of the implementation and the
3 The HFST-TWOLC Grammar **In the following, I will go through** the two-level grammars for mapping the

4 from the same proto-language**, I will first go through** a more conventional example of the use
5 the sound changes individual cognates **go through** while diverging from the proto-
6 are characterized by having a base, which may **go through** different variation, and a complicated

(Kaisa, C1, concordances for the concgram GO/THROUGH)

Both concgrams strongly associate with the function of metatextual commenting and participate in the following pattern: reference to the text (*in the following, this section, in this section*) +/or reference to the author (*I*) + a discourse label in the future tense naming the discourse act to be performed (*will go through X, will concentrate on demonstrating/describing/reconstructing*). The construction of such a pattern will be discussed in more detail in the next chapter in Section 6.3.3 together with the analysis of a similar pattern which is also a remarkable example of fixing: *In the following I will* + verb. Here it is enough to point out that metatext seems to be a fertile ground for fixing because certain functions are repeatedly required to be verbalised throughout the text or, in other words, certain discourse acts need to be performed over and over again. Therefore, through repetition a preference for a certain wording a language user develops becomes fixed.

The examples presented in this section were indeed able to illustrate the routes through which fixing can work. They showed how units of meaning acquire new components (a collocation of *cognate recognition* with *automatic*, or a semantic preference of the verb POSTULATE for 'sounds'), how already existing components change their status from a more abstract association like semantic preference to a collocation (the case of *cognate recognition* becoming preferred over *cognate identification* and *cognate detection*) or the lexical filling of a metatextual frame becomes fixed. In all, fixing appears to be a normal process in language: it prepares ground for the meaning-shift and emergence of fixed idiomatic expressions which become opaque and non-compositional in the end (see also discussion in Section 2.4).

In all, approximation and fixing seem to be the candidates for the processes which underlie operation on the idiom principle. They seem to drive variation and change in units of meaning without making the production mechanism switch from idiom principle to open-choice.

One more observation seems to be in order. Together with Matching but 'overused' patterns and individual preferences, patterns of approximation and fixing show that abrupt changes between the priming language and the language produced are rare. For example one would be inclined to regard the case of *antiretroviral treatment* in C1 vs. *antiretroviral therapy* in C2 as an approximation unless two examples of *antiretroviral treatment* were found in C2. Yet, since they were found, the use of *antiretroviral treatment* instead of *antiretroviral therapy* is just a preferred usage on the part of the writer which is also becoming fixed due to repeated demand for use. Also, it is just one of the two alternatives that she uses: this is an example of how a preference becomes instantly fixed. This is also one of the major reasons why it was so hard to find approximated associations, semantic preference and colligation, in the data. When a unit is produced for the first time, it might be a semantic preference or a colligation working, but already starting from the second use, it is a verbatim association, i.e. a collocation, which is at play.

5.6. Conclusions

This chapter set as its goal to examine the degree to which the idiom principle is available to L2 users both in language acquisition and use. Operation on the idiom principle was probed by different means.

First we looked at isolated patterns occurring in the data to see whether they represent recurring units of meaning, and therefore demonstrate operation on the idiom principle, using the following criteria: (1) recurrence, (2) compliance to the model of a unit of meaning and (3) consistency of meaning communicated.

Second, since the corpora of expert writing were compiled from the articles students referred to in their theses, with the aim of representing their priming language, it was feasible to compare significant 2-word co-occurrences from students' writing to the patterning of the expert writing. The results showed that more than half of the patterns students use match the patterns of the expert writing in their field, and therefore at least some of them are likely to have been acquired directly from this writing. These results imply that the patterns students use were not constructed in the first place but acquired holistically from exposure suggesting that L2 users are able to acquire lexico-grammatical patterns on the idiom principle too.

Finally, the students' lists of significant 2-word co-occurrences, both matching expert writing patterning and not matching it, were analysed qualitatively using concordance data from both students' corpora and expert writing corpora. The analysis showed that matching patterns correspond to the field-specific patterning exhibited in expert writing down to finer details, such as additional field-specific components in otherwise common units of meaning, distributional patterning and even some patterning which conflicts with more commonly used alternatives like in the example of *remotely* vs. *distantly related.* A lot of the patterns which did not match expert writing were content or genre specific. The remaining non-matching co-occurrences seemed to reveal two processes: approximation and fixing.

The two are reverse processes. Through approximation a relatively established pattern becomes more relaxed as some variability is introduced to it. But the association between different components of a unit of meaning can also become stronger. As a result, they become more fixed: some optional, vaguely present components can become more habitual and other more abstract components can start to co-occur verbatim. This process of fixing seems to be the initial point for delexicalisation and meaning-shift which may follow in case the pattern continues to be repeatedly used and its communicative purpose becomes more and more established. All of this strongly suggests that approximation and fixing are normal processes accompanying operation on the idiom principle, or, in other words, they are part of the idiom principle. In Sections 6.2.1 and 6.3.9, it will be shown that both approximation and fixing can be observed in word association responses too, which indicates that they reflect the processes representations of lexico-grammatical patterns undergo in the mind.

In sum, we have seen that the patterns L2 users employ resemble units of meaning which are produced on the idiom principle. These patterns tend to come directly from the language L2 users were exposed to. The patterns which depart from the patterns common to expert writing seem to arise either as a result of content and genre specificity, or through the processes of approximation and fixing which, as has been argued, are aspects of operation on the idiom principle. In other words, there is nothing in the data which speaks for the tendency of second language users to operate on the open-choice principle. Above all, we can conclude that L2 users are sensitive to the phraseological tendency of language and tend to organise their language syntagmatically as well. It seems that the idiom principle is available to L2 users to a much larger degree than is usually claimed. They are not only using lexical patterns on the idiom principle but also acquiring them implicitly from exposure, i.e. on the idiom principle.

In the next chapter, I will compare C1 patterns with word association responses to examine whether the units L2 users seemingly acquire and produce on the idiom principle are also holistically represented in the mind. In particular, in Sections 6.3.8 and 6.3.9, the evidence from both comparisons, C1 vs. C2 and C1 vs. WA responses, will be brought together to find out whether there is continuity between the patterns in the priming data, the patterns produced and the patterns produced by syntagmatic association in WATs.

6. The psycholinguistic reality of a unit of meaning: C1 vs. WA responses

Chapter 5 compared L2 usage patterns to the priming language and showed that they are not only produced on the idiom principle but are also likely to be learned implicitly from exposure, i.e. on the idiom principle. In this chapter, I will compare L2 usage patterns with word association (WA) responses to see whether there is evidence that these usage patterns are also holistically represented in the mind, i.e. can be processed on the idiom principle.

First, I will ascertain whether there is a relationship between the usage patterns a word participates in and WA responses it elicits in a word association task (WAT). Then I will focus on syntagmatic WAs and investigate whether this syntagmatic association can underpin operation on the idiom principle. I will also explore this syntagmatic association further. The main goal of this chapter is to examine to what extent evidence can be found for the psycholinguistic reality of the model of a unit of meaning.[72] More specifically, it is intended to find out whether the components of a unit – collocation, colligation and semantic preference – can be represented in the mind, or, in other words, whether syntagmatic association underlying the idiom principle can both be verbatim and abstracted semantically or grammatically.

Methodologically (see Ch. 4, Section 4.2.4 in particular), the analysis in this chapter is structured similarly to the analysis of Chapter 5. While in Chapter 5 C1 usage patterns were compared to the priming language represented by C2 corpora, in this chapter C1 usage patterns will be compared to the students' WA responses. Thus, just as in Chapter 5, we will be dealing with Matching and Non-matching patterns or in this case Matching and Non-matching WA responses. In contrast to usage patterns, WA responses can be not only syntagmatic but also meaning-based or paradigmatic, which by definition cannot match usage patterns. Therefore, in order to compare WA responses to usage patterns, it is first necessary to classify them in their own right, which will be done in Section 6.1. In Section 6.2, WA responses are further categorised based on how they compare to the usage patterns. Section 6.3 capitalises on the observations made in this categorisation and develops them further

[72] This is not the first attempt at exploring the psycholinguistic reality of the components of Sinclair's unit of meaning. Ellis et al. (2009) and Ellis and Frey (2009) looked at the psycholinguistic reality of collocation and semantic prosody at the stages of word recognition and lexical access and at the stage of semantic access, respectively. However, the definitions of collocation and semantic prosody they used in their experiments were different from the conceptualisation of this study. Collocation was defined as a "co-occurrence of particular words" and semantic prosody as the phenomenon "whereby a word can be associated with generalized types of words, for example verbs with negative rather than positive objects" (Ellis et al. 2009: 94-95). In this study, collocation and semantic prosody are first and foremost seen as components of a unit of meaning with the underlying conceptualisation of lexis and meaning.

focusing on each specific effect or tendency noticed. In the last two subsections, 6.3.8 and 6.3.9, the comparisons of usage patterns with WA responses and with the priming language will be brought together. A summary of the conclusions is provided in Section 6.4.

6.1. Classification of WA responses

To compare WA responses to usage patterns, it is necessary to classify them first. As a rule, in WA studies responses are classified into syntagmatic or position-based, paradigmatic or meaning-based and clang or form-based (see Section 4.2.5 for a review of WA studies). The present study adopts this classification framework, but the categories are interpreted differently. The major difference from the previous studies is the connection with the idiom principle which is tested here.

The subsections that follow focus on meaning-based and syntagmatic responses only. The reason why form-based responses are not taken up is that they are extremely rare in my data. This is not particularly surprising as in previous studies it was noticed that form-based responses, i.e. responses which concentrate on the formal features of the word, are more typical of earlier stages of language acquisition both in L1 (see Fitzpatrick 2007: 321) and L2 (e.g. Meara 1983) with adult native speakers hardly ever responding with form-based associations at all. Also, the criteria for form-based responses I used are even narrower than in previous studies. For example, Fitzpatrick (2007, 2009) distinguishes between form-based responses which are only similar in form to the stimulus word, not meaning, like simple clang responses, and responses which involve a change of affix. In this study, only the responses of the first type were classified as form-based. The responses *semi-structured → structured* (Linda); *possible → impossible* (Kaisa) and *efficient → inefficient* (Kaisa) are all interpreted as meaning-based. Even though only a change of an affix makes them different from their stimulus words, there is a semantic relation between a stimulus and a response, that of antonymy. Thus, it is possible to assume that the respondents had to access the meaning of the stimuli to produce their responses. With this adjustment, the WATs that were selected for analysis contained only one response which could be indeed purely form-based: *other → neither* (Linda) – it was excluded from further analysis.[73]

[73] It is not the only response which was excluded from analysis. There were cases when a chaining or interference effect was apparent, for example when six stimuli were responded with the same association, *done*, even though in each of the cases it was a reasonable response which the respondent was able to explain. Also, some responses did not have enough complementing data to inform the classification. They were excluded too. The exact number s of responses which were analysed are provided in Table 6.9.

So, in the following subsections (6.1.1 - 6.1.2), I will explain the grounds on which I based the classification of WA responses into meaning-based and syntagmatic. I use different means to describe each group for several reasons. For the explanation of meaning-based responses, the retrospective comments that respondents give play a major role because they clarify the relationship that the respondent sees between a stimulus word and a response. Therefore, I will try to give as many examples of associations and retrospective comments as possible. Also, it seems helpful to divide meaning-based responses into several groups in order to see what unites them, a procedure which needs a good number of examples as well. At the same time, the sub-classification of meaning-based responses which is provided in Section 6.1.1 is only used as a method of description and is not claimed to be exhaustive or consist of clear-cut groups. In contrast, syntagmatic responses are best elucidated by showing their interrelationship with usage: students' own usage patterns or the patterning of general language use as represented by general purpose corpora. Therefore, due to the extended analysis which is required in this case, the number of examples has to be limited. Yet, the small number of examples is not a problem, since later on in the chapter, the focus will shift to syntagmatic responses and there will be many examples provided when WA responses will be compared to usage patterns in Section 6.2.

In this section, I will not give the raw frequencies of the different kinds of responses. As pointed out in Section 4.2.5, a response which is produced in a WAT is influenced not only by the associations internalised by the respondent but also by the properties of the stimulus word itself. Since the stimuli were selected randomly, the frequencies of different types of responses make sense only when the properties of the stimuli are revealed, that is, when WAs are compared to usage patterns, which will be done in Section 6.2, with the results of the quantitative comparison presented in Section 6.3.6.

6.1.1. Meaning-based (M) responses

In this subsection, I describe meaning-based responses to give an overview of the responses produced.

One stimulus-response pair can be categorised in more than one way. For example, the stimulus-response pair *hamlet* → *village* is analysed as an example of hyponymy, but it could also be synonymy which underlies the relationship, or the association *somewhat* → *certain* is suggested to be rooted in the connotation of *somewhat*, but it can also be described as an example of antonymy. In other words, the purpose of the description is not to provide a watertight classification, but to demonstrate the meaning-based nature of this type of association. The responses seem to be arrived by inference rather than by remembering. The impression is that the respondents see a stimulus word as a meaningful entity and try to interpret or explain its meaning. The ways in which they do it are remarkably similar to structuralist analysis of semantic relations between words. Therefore, it seemed natural to describe meaning-based responses in terms of the semantic relations they designate.

I will start with responses which seem to illustrate conventional semantic relations described in the field of lexical semantics and then move on to fuzzier cases. Examples of the commonly acknowledged semantic relations which were observed in WA responses are antonymy (6.1), synonymy, including partial synonymy or sense synonymy when stimulus and response words are synonymous in a particular sense only (6.2), meronymy (6.3), hyponymy (6.4), hypernymy (6.5), co-hyponymy (6.6).

(6.1)

Antonymy:
automatically → *by hand* ("antonyms", Kaisa);[74]
seldom → *often* ("antonyms", Kaisa);
woman → *man* ("paired with man, opposite", Hertta);
beginning → *end* ("opposite in some way", Hertta);
civil → *military* ("makes me think the opposite", Maisa);
decline → *increase* ("these are the words that they use really very much in there", Maisa).

(6.2)

Synonymy:
enables → *possible* ("if you enable something, then you make it possible", Kaisa);
aware → *awake* ("if you are awake, you are aware, if you are asleep, you are not", Kaisa);
appears → *seems* ("synonyms", Kaisa);
crucial → *important* (Kaisa*);*
distributed → *spread* ("if you distribute something then you spread it out", Kaisa);
interpret → *read* ("if you interpret something then that is your reading of the thing", Kaisa);
contribute → *give* ("to give something to somebody", Maisa);
nevertheless → *however* ("kind of synonyms", Linda);
indicate → *show* ("indicate is a little bit more specific than shows",[75] Kaisa);
prevent → *stop* ("feels the same", Hertta);
argued → *said* ("he said something, argued is stronger though", Linda);
affected → *effect* ("just the same kind of word, maybe I use them both", Maisa);

[74] In parenthesis I will cite retrospective comments of the respondent in question: after taking a WAT, she was asked to briefly explain why she thought she had given each of the responses. These responses are very helpful in informing the categorisation of WA responses.
[75] Quite clearly, the respondent is referring to the academic context here.

term → *concept* (“I was playing with those two”, [76] Linda);
ways → *means* (“there were communication ways and means, I was playing with those words [in the thesis]”, [77] Linda).

(6.3)

Meronymy:

code → *program* (“program consists of codes”, Kaisa);
households → *demography* (“thought of this as a demographic concept”,[78] Maisa).

(6.4)

Hyponymy:

vertebra → *bone* (“vertebra is a bone”, Hertta);
hamlet → *village* (“because hamlet is a smaller village”, Hertta);
ungulates → *animals* (“they are large mammals specific time animals”, Hertta).

(6.5)

Hypernymy:

constructions → *buildings* (“buildings are some sort of constructions”, Hertta);
institutional → *hospital* (“or I could have said prison or something”,[79] Maisa).

(6.6)

Co-hyponymy:

axes → *knives* (“there is axes in the graves sometimes and there is also knives in the graves sometimes”, Hertta);[80]
child → *infant* (“…it is sort of at least in the same age group”, Hertta);
assumptions → *ideas* (“something going around in peoples’ heads, perceptions [..] ideas, assumptions”, Linda).

Other responses seem to need a bit more explanation. For example, there are responses which look as if the respondents were deconstructing the meaning of a stimulus word and identifying one of its components, semes, imitating the structuralist idea of componential analysis, as in (6.7)

[76] It is a relatively common comment that the stimulus and its response are interchangeable in the context of their thesis.

[77] Synonyms by collocational behaviour.

[78] So the concept of a household belongs to the field of demography.

[79] As it transpires form Maisa’s reference corpus, hospital and prison are regarded as types of institutional households.

[80] This is an example of a very contextualised type of co-hyponymy because the hypernym could be termed something like ‘items which are usually found in the graves’.

(6.7)

expanded → *bigger* ("if you expand something it usually gets bigger", Kaisa);[81]
consequence → *follows* ("something follows from that", Kaisa);
contribution → *add* ("if you make a contribution, you add", Kaisa);
covered → *layer* ("if something is covered, it has a layer", Kaisa);
depending → *relation* ("if you depend on something then there is some sort of a relation there", Kaisa);
authority → *leader* ("authority and then you are leader", Linda);
interpersonal → *together* ("interpersonal: you do something together with people", Linda);
previous → *before* ("before something", Linda).

In other responses (6.8), one can see the respondent trying to characterise connotation of a stimulus word.

(6.8)

obvious → *stupid* ("cause if you say isn't that obvious then you sort of imply are you that stupid that you can't see that that is obvious", Kaisa);
required → *must* (Kaisa);
mandatory → *must* (Kaisa);
inevitable → *must*[82] ("then it's a must", Kaisa);
discard → *bad* ("if you have like bad forms or whatever you have to discard it from your data", Kaisa);[83]
beneficial → *good* ("beneficial is always good", Linda);
relevant → *important* ("relevant, it is usually important", Linda);
somewhat → *certain* ("that was hard, I guess I thought certain would be the opposite, somewhat there, somewhat here, when you are certain it is not somewhat anymore", Linda)[84].

Another group of responses, illustrated in (6.9), is in a way very similar to the previous one, yet, instead of characterising the connotation of the stimulus word, the response seems to

[81] *expand*: "become or make larger or more extensive" (*Oxford Dictionary of English* 2010)

[82] These "must"-responses come from the same student but in different WATs with approximately three and four months in between them (*required* - WAT2, item N 33, 14.09.09, *mandatory* - WAT4, item N62, 02.12.09, *inevitable* - WAT5, item N8, 13.04.10, item numbers are important in controlling for 'interference' effects which will be given more attention later on in this chapter).

[83] Though the respondent is interpreting the meaning which is contained in the stimulus word alone, some responses are more contextualised than others.

[84] While the explanation given by the student seems to demonstrate that it was a meaning-based response, it could have been syntagmatic. *Uncertain* is the 20th most frequent collocate of s*omewhat* in the position 1 to the right in the BNC. Perhaps the response given was based on a syntagmatic association in the first place which was subconscious. It was then rationalised and turned to *certain.* The respondent's comment might be describing either the rationalisation process or a post hoc explanation of the association she did not really know how she got.

explain its implication or perhaps the consequences that may follow from the meaning of the concept associated with the stimulus word

(6.9)

efficiently → *fast* ("sometimes it means that if it is efficient, it is fast", Kaisa);
consistent → *sure* ("I guess, you are sure of something if you are consistent", Linda);
focus → *highlighting* ("focus, you highlight something", Linda);
reluctant → *want* ("you don't want to do something when you are reluctant", Linda).

In yet other responses, the respondents seem to be actually projecting a stimulus word onto its referent in the real word and describing what this would entail, i.e. they are acting out the meaning of a word in their imagination (6.10).

(6.10)

evoke → *push* ("well if you evoke something then you sort of try to wake it up, and that you can do by pushing", Kaisa);
replaces → *removes* ("…then you remove the first thing and insert the second thing", Kaisa);
ignore → *skip* ("if you ignore something, you sort of skip ahead", Kaisa);
reveal → *cover* ("it was more about the actual action, you reveal something, it has been covered before", Linda).

Example (6.11) shows responses which are a bit more abstract than synonymous responses and can be regarded as belonging to the same lexical field as their stimuli.

(6.11)

exemplify → *instance* ("if you exemplify then you give an instance of something", Kaisa);
communicating → *speak* ("you communicate when you speak", Linda);
skills → *talent* (Linda);
antenatal → *pregnancy* ("here is a lot of stuff about pregnant women, so", Maisa);
design → *fashion* ("I know I am talking about design here [in the thesis] and it's really a different thing, it is funny you had here trends, and I was also thinking about clothes [N99 *trends* → *clothes*]", Maisa).

This analysis leads to the conclusion that in meaning-based associations we can see semantic connections which we usually use in glossing the meaning of lexical items, for example in dictionaries or theories of lexical semantics (see e.g. Geeraerts 2010). Most importantly it can be claimed that when a respondent gives a meaning-based response, she responds to concepts or meanings, or components of them. So, such responses can also be described as word-internal. Therefore, it seems that to elicit a meaning-based response, a stimulus word needs to have a meaning which is interpretable, i.e. a meaning which is relatively complete without the help of other words.[85] It also seems possible to hypothesise that meaning-based responses are usually a result of declarative processing. This hypothesis will be further discussed in Section 6.3.7.

6.1.2. Syntagmatic (S) responses

In contrast to an M-response, a syntagmatic response or an S-response does not involve interpretation or paradigmatic choices. It can be said that an S-response is in a way produced on the idiom principle. In this study, an S-response does not have to correspond to an established word combination: a Standard English collocation or an idiom. It is sufficient that a stimulus word and its response can be used together meaningfully, i.e. participate in a unit of meaning contributing to its semantic prosody.

Therefore, not only WAs like *comparative → method* (Kaisa); *computational → linguistics* (Kaisa); *underground → culture* (Nora); *based → on* (Hertta); or *take → into account* are classified as S-responses, but also associations like *task → hard* (Linda); *strong → light* (Linda); *for → me* (Kaisa); *access → Microsoft* (Maisa, [Microsoft Access is a programme from Microsoft]) or *registered → births* (Maisa). All these and similar examples will be discussed in Section 6.2.

At the same time if only those responses which conform to the model of a unit of meaning were classified as syntagmatic, the hypothesis that it is the units of meaning which are reproduced in WATs would become untestable. In particular, this would not allow for a possibility of a syntagmatic association outside the boundaries of a unit of meaning. Therefore, in cases where a stimulus word and its response are used positionally close to each other without forming a clear shared meaning, that is, the meaning of the pattern does not appear to have evolved into a separate communicative function but is still largely compositional, the response is still categorised as syntagmatic.

For example, Kaisa had an association *perceived → similar* supplying it with a comment “as a phrase”. However, it does not seem to be a phrase, and indeed a search in the BNC of the lemma PERCEIVE accompanied by *similar* in the span four to the right gives only

[85] As pointed out in Section 2.8.2, it is not necessarily a word which is needed to make the patterning of a unit of meaning complete, in agglutinative languages it can be a case ending, but in English we are usually bound to words. Yet, as Mauranen (2012:101-102) shows, approximation can occur inside a single word too, like in the examples of *successing* instead of *succeeding* and *negated* instead of *denied*. These means that morphological elements can also be treated as meaningful components and thus get approximated. So “saying without the help of other words” is a shortcut which should be read “without the help of other linguistic elements”, to be exact.

seven quite random hits. At the same time in Kaisa's own texts *perceived as similar* is used four times, as can be seen from Example (6.12).

(6.12)

1 words are cognates if they are **perceived as similar** and if they are translation kaisa_10_300410.txt
2 dictionary. False friends are **perceived as similar** but have different meanings kaisa_10_300410.txt
3 words are cognates if they are **perceived as similar** and if they are translation kaisa_summary_210109.txt
4 dictionary. False friends are **perceived as similar** but have different meaning kaisa_summary_210109.txt

(Kaisa, C1, concordances for the lemma PERCEIVE)

The concordances in (6.12) suggest that the use of *perceived as similar* is consistent and associates with one meaning. Yet, it is too early to say that the pattern is developing into a unit of meaning with a structure of a collocation, for example, because it is used in only one context, and it is difficult to think of other possible contexts where this word combination can function as a unit. At the same time, the frequency of four is considerable for a corpus of somebody's drafts, even though two of the uses are mere repetitions. An extended context view (6.13) shows that the two uses actually occur in the adjoining sentences in the context of defining cognates and false friends.

(6.13)

According to the authors, words are cognates if they are perceived as similar and if they are translations of each other in a bilingual dictionary. False friends are perceived as similar but have different meanings.

(Kaisa, C1)

In giving the definition, Kaisa refers to an article she used for her summary, and concordances from her reference corpus (6.14) show that she has probably borrowed this particular phrasing from there.

(6.14)

1 riends (Vrais Amis), are pairs of words that are **perceived as similar** and are mutual translations. The spell
2 are pairs of words in two languages that are **perceived as similar** but have different meanings, e. g.,

(Kaisa, C2, concordances for the lemma PERCEIVE)

What is important here is that as is evident from Kaisa's WA response, the pattern probably becomes represented in her memory as a whole. The fact that *similar* is elicited in response to *perceived* indicates that the two words become syntagmatically associated, and therefore, the response is categorised as syntagmatic. Possibly, this is evidence of the process of fixing suggested in Chapter 5. It is also possible that in favourable contexts the pattern can develop into a unit of meaning for this student in case she finds other applications for this syntagmatic association she already seems to have.

As the analysis of the phrase *perceived as similar* shows, in order to determine whether a response is syntagmatic, different types of evidence were used: the respondent's retrospective comments, concordance data from the respondent's writing samples, concordance data from the respondent's reference corpus. Yet, there is one more type of evidence which was not yet brought to the fore: just like in the previous chapter, in cases where a stimulus-response pair seemed to relate to general rather than field-specific English, it was compared to the patterns represented in general purpose corpora, the BNC or sometimes also the COCA. For example, Hertta produced a stimulus-response pair *opposite → direction*. Her comment: "there is an opposite direction maybe somewhere" implies that she regards the word combination as a phrase, but there is only one occurrence of *opposite direction* in her writing samples, even though the total frequency of *opposite* is not much higher (n=3). At the same time, in the BNC, *direction* is the second most frequent collocate of *opposite* after the definite article *the.* Therefore, taking all the evidence into account, *direction* is categorised as an S-response.

It is not always easy to distinguish a syntagmatic response from a meaning-based one. A response which at first glance appears to be syntagmatic since it can be put together with a stimulus word in a syntagm may in fact be used to open up the meaning of a stimulus word. Precisely in the Firthian sense of the modes of meaning, by providing the context for the stimulus, the response may be specifying its meaning. Here are some examples from Kaisa's WATs: *architecture → house* with "well, architects build houses" as an explanation or *bears → forest*: "bears live in the forest". The respondent has managed to put the two words together in a sequence, but it does not yet mean that they form a unit of meaning. More likely, they bring to mind Firth's example "cows give milk" as a way of distinguishing cows from tigresses. This line of reasoning is supported by corpus searches.

In the COCA, HOUSE is not among 100 most frequent collocates of ARCHITECT, and neither is *forest* among 100 most frequent collocates of the plural or singular for the noun *bear.* So the decision is to assign these responses to M, meaning-based. In contrast, some very similar responses like *treated → doctor* ("doctor treats patients"); *reflect → mirror; responsible → adult; vocabulary → learn* ("to learn vocabulary") or *released → album,* which intuitively qualify better to be categorised as syntagmatic, present a different picture. All of these responses are among the most frequent collocates of their respective stimulus words. *Adult* and *adults* are the 10th and 11th most frequent collocates of *responsible* in the position one to the right, *album* is the sixth most frequent collocate of RELEASE *(albums* – 30th*), mirror* is the third most frequent collocate of REFLECT (*mirrors* – 21st), *doctors* is the 12th most frequent collocate of TREAT and *learn* is the 15th most frequent collocate of *vocabulary* in the span four to the right and four to the left. A bit different case is *create* - *God* ("God created"). *God* is not among the most frequent collocates of CREATE, but *creation* and *creator* are among the most frequent collocates of *God.* This response is also classified as S.

While M-responses are arrived at by inference, syntagmatic responses or S-responses are assumed to be more spontaneous: in order to give an S-response one does not presumably have to access the meaning of a stimulus word and interpret it but just rely on simple retrieval. In the last several decades of research, it has often been stated that language production by remembering requires much less processing effort than language production by constructing from scratch on the basis of the rules of grammar (e.g. Bolinger 1976; Ellis 1996; Wray 2002). So if there is something to remember, then the natural tendency would be to fall on these resources. It is a common observation that in repetitive tasks like giving the same lecture or the same guided tour,[86] the speakers often find themselves saying the same things wrapped up in the same words (see Peters 1983: 8; Wray 2002: 5 for similar observations). The WAT respondents in this study have these resources: they have produced the texts comprising C1 corpora in a high-stakes context rather than just experienced them receptively. Since these are the drafts of their Master's theses, they have spent a considerable time on conceptualising, writing and editing them. So it seems reasonable to expect that they would resort to these texts in their memory when giving responses to a WAT where they are encouraged to respond as fast as possible. The routine and tedious nature of the task which was given at every meeting, amounting to four-five tasks per student on average, with each of the tasks involving giving responses to at least a hundred stimuli in one go, were also conducive to resorting to the easiest way of handling the task from the processing point of view.

So in contrast to M-responses, which are assumed to be given by inference, S-responses are thought to be given by retrieval from implicit memory, that is, on the idiom principle. An M-response will therefore be elicited by a stimulus word with a relatively

[86] In my own experience of giving guided tours, quite soon the amount of practice becomes a problem rather than an aid: one feels trapped in the text, and it becomes very hard to get off the beaten track of habitually used chunks and structures.

complete meaning, a word whose use is independent. In contrast an S-response will be elicited by a stimulus word which does not have a complete meaning, a word which is delexical and dependent in use. The hypothesis will be explored by comparing WA responses and usage pattern in Section 6.2, and the significance of the interrelationship will be tested in Section 6.3.6.

Another important observation is that respondents of WATs seek to make their responses meaningful: form-based responses are rare, in M-responses the meaning is attributed to the stimulus word itself and in S-responses a new unit of meaning is brought up by completion of the pattern the stimulus word forms. Most of the time the respondents of this study are able to provide a reasonable explanation of their association. Sometimes they say that they do not know why they have given such a response because they do not remember or because the association they have provided no longer makes sense to them, but they never say that the response is given for no particular reason.

6.2. Comparing WA responses to C1 patterns

So, in the first stage of the analysis, it was decided whether a stimulus word elicits a syntagmatic (S) or meaning-based response (M). In the second stage this response is compared to C1 usage patterns the stimulus word participates in. The corpus can reveal that a stimulus word is in dependent or independent use.[87] That is, a stimulus word can be used as part of an extended unit of meaning (MWU) or as a self-contained lexical item with a relatively complete meaning of its own without being dependent on the accompanying words (No MWU). As explained in the previous section using the example of the pattern *perceived as similar*, not all multi-word patterns which can be detected in C1 can be unhesitatingly called units of meaning. Also, making an a priori assumption that all syntagmatic patterns are necessarily units of meaning does not permit testing the possibility that syntagmatic association can exist outside a unit of meaning (this possibility will be discussed in Section 6.3.2). For these reasons, the first scenario, when C1 reveals a syntagmatic pattern for a stimulus word, is called MWU-scenario, rather than unit of meaning or MSU-scenario, and the second, when there is no pattern identified – No MWU-scenario.

[87] The terms independent and dependent are suggested by Sinclair. For example he writes: "Intuitively, we feel that some instances of a word are quite independently chosen, while in other case we feel that the word combines with others to deliver a single multi-word unit of meaning. We shall call word-meaning *independent*, and phrase-meaning *dependent"* (Sinclair 1991: 71). So we cannot say that a word can be dependent or independent, what can be dependent and independent is its use.

When WA responses are compared to usage patterns, a WA response can match a MWU found in C1 (Matching MWU-scenario), it may not match it (Non-matching MWU-scenario), or it may turn out that C1 does not show any pattern to compare a WA response with, that is a stimulus word is independent in usage (No MWU). The cross-tabulation of syntagmatic and meaning-based categories with Matching MWU, Non-matching MWU and No MWU categories results in five major categories, which will be described and discussed in detail in the following sections: (1) Matching MWU S-responses, (2) Non-matching MWU S-responses, (3) Non-matching MWU M-responses, (4) No MWU S-responses, (5) No MWU M-responses. The cross-tabulation is presented in Table 6.1. The table provides examples of each category by giving a WA stimulus-response pair and, in MWU-scenarios, a corresponding MWU a stimulus word participates in.

Table 6.1 Comparing WA responses vs. C1 patterns

Response category	Syntagmatic responses	Meaning-based responses
Matching MWU	*according → to : according to* *comparative → method : comparative method* *splitting → two : (BE) split into two/ SPLIT into two*	
Non-matching MWU	*makes → sense : This* (noun) *makes* + noun + adj. *bilingual → children : bilingual data* *for → me : for the sake of, for instance*	*dealing → handling: DEAL with* *method → way: comparative method*
No MWU	*appendix → burst* *actually → love*	*also → plus* *whereas → different* *inevitable → must* *laborious → stupid*

There is no Matching M-response category because M-responses which would match MWU do not exist. A response would not even be categorised as meaning-based if it showed a co-occurrence relationship with the stimulus-word.

The presentation of categories in the subsection will follow the order set by Table 6.1.

6.2.1. Matching MWU S-responses

A Matching MWU S-response is a syntagmatic response which also matches a MWU used in the text. For example, *projection* elicits *population* ("population projection", Maisa) and this stimulus-response pair also has a corresponding bigram *population projection* which occurs 20 times[88] in the respondent's C1. Therefore, the response *population,* which was categorised as an S-response in the first stage, is also categorised as Matching MWU.

A lot of Matching MWU S-responses can be described as field-specific terminology. Examples of such responses elicited from different respondents are provided in (6.15).

(6.15)

Kaisa: *TWiki → platform; comparative → method; computational → linguistics; machine → translation; optimality → theory; orthographic → similarity; proto → language; similarity → measure; training → data; initial → sound; version → control; surface → realization; general → linguistics; automatic → recognition (automatic cognate recognition).*

Hertta: *inhumation → burial*; *large → mammal* ("I have large mammals in the texts, if I cannot identify them into species they are just large mammals"); *trial → excavation* ("that was a word that [...] supervisor suggested because I had written test and she marked that it should be a trial and after that I have sort of noticed that when I read texts that it is a trial excavation, funny"); *goods → grave* (*grave goods*).

Linda: *affairs → foreign* (*foreign affairs*); *communication → activities; conferences → press* (*press conferences*); *decision → maker; diplomacy → public* (*public diplomacy*); *federal → chancellor; target → group.*

Maisa: *census → data (census data); sentinel → survey (sentinel survey); spectrum → software (Spectrum software); projection → population (population projection); structure → population (population structure); errors → sampling (sampling errors).*

Nora: *authentic → identity; mainstream → culture; heavy → metal; shared → space; online → research (online research methods); underground → culture (underground subculture).*

Other Matching MWU S-responses represent more 'general' English patterns, as in (6.16).

[88] The frequency of occurrence does not play a major role here because the fact of occurrence itself together with a Matching syntagmatic response is enough to establish a syntagmatic association: it is hardly accidental that two words co-occur in usage and elicit each other in a WAT. At the same time it is rare that a decision to categorise an S-response as Matching MWU is based on one occurrence only, like in the case of Kaisa's association *relatedness → language* where *language relatedness* occurs just once in her C1 but her own comment on the WA response is "from the text". So while in some cases I will provide information on the frequency of occurrence, like here, to show, for example, how strong or weak a co-occurrence relationship is in general terms, I will not do it systematically.

(6.16)

Hertta: *based → on; derive → from; situated → in.*

Kaisa: *look → at; rely → on; benefits → of; caused → by; concentrate → on; descend → from; each → other; even → though; refer → to; results → from; take → into account; use → of.*

Linda: *access → to; despite → fact* (*despite the fact*); *kind → of; reasons → why.*

Maisa: *derived → from; instance → for.*

Nora: *revolves → around; access → to; aware → of; due → to; every → body; everyday → life; refer → to; arise → from; dealing → with; oppose → to; amount → of something.*

These first two groups of Matching MWU S-responses reflect relatively fixed patterns, either general English or field-specific, but some associations, while often verbatim for the respondent, would be considered looser from the point of view of English as a whole. Possibly this is how a semantic preference is shaped: while for each individual language speaker a certain word collocates with a certain other word, when taken together, the different collocations form a preference for a semantic set. Let us look at some of such associations that are less fixed but collocating for the test taker in Example (6.17).

(6.17)

Kaisa: *closely → related (closely related languages); significant → difference; systematic → change; perceived → similar (perceived as similar); solution → problem (solution to the problem).*

Hertta: *thick → layer* ("it's a thick layer of earth or something"); *damaged → graves* ("I know that I use damaged when I talk about graves because they are sort of partially damaged, people have built or excavated before or sort of made newer graves on top of the older ones"); *earth → mixed* (*earth mixed with stones*; 4 occurrences in C1).

Maisa: *methods → section; overnight → travellers.*

Nora: *clearly → stated; existing → research.*

Linda: *driven → mass media* (*mass media driven*); *time → lack* (*lack of time*)[89].

Since, out of all S-responses, only Matching MWU S-responses can be compared to corpus data from the respondent's use, only these responses can be categorised in terms of their syntagmatic association into collocation, colligation and semantic preference. All of these are

[89] Both of the phrases are used just once in her corpus, perhaps it is the recency factor which affected her WA responses.

syntagmatic associations which work within a unit of meaning, in other words all responses assigned to these categories are also unit of meaning responses. However, as I have already mentioned, it seems that syntagmatic association is also possible outside a unit of meaning, and these cases will be discussed in Section 6.3.2. Yet, the vast majority of Matching MWU S-responses can be interpreted within the concept of a unit of meaning.

All the examples discussed in this section so far are collocational demonstrating a verbatim association between words. Let us look at some of them more closely.

Linda associates *reasons* with *why,* the corresponding collocation in her corpus is *the (*) reasons why,* as (6.18) shows.

(6.18)

1 Merkel. After a brief presentation of the **reason**s **why** a qualitative methods and semi-structured
2 in the first section I discuss the general **reason**s **why** they do not manage her international image

(Linda, C1, concordances for the concgram REASON*/WHY)

In the same way she associates *decision* with *maker* in a WAT, and accordingly she is writing about *decision makers* in her texts (6.19).

(6.19)

1 the latter has only little affect on the elite **decision makers** [ref.]. However, this may change
2 to present her as a modern, effective **decision maker**. It is important to present her a
3 professors, non-governmental organizations, **decision makers** and leaders from private industry.

(Linda, C1, concordances for the concgram DECISION/MAKER*)

Reasons why and *decision maker* are examples of contiguous collocations where one word follows the other without any intervening words. But non-contiguous collocations are also represented in the WATs. For example, Hertta associates *thick* with *layer* commenting “it’s a thick layer of earth or something”. The concordance for the concgram THICK/LAYER extracted from her corpus (6.20) shows that the association between the two words is less precise than in the previous examples.

(6.20)

1 layer of sooty soil was found in a 3-4 cm **thick layer**. Under the second sooty soil, about 0,3 cm
2 to 100 cm but with an average of 35-40 cm. The **thickest layer**s were presumably produced by trash
3 unfurnished. The partly unused lands had a bit **thicker** soil **layer**s reaching up to 100 cm but with an
4 sand [ref.]. From the **thickness** of the cultural **layer**s [ref.]
5 where the surface **layers** were considerably **thicker** than in the other areas. The base of the pits

(Hertta, C1, concordances for the concgram THICK*/LAYER)

Contiguous collocations can be further divided into XY and YX collocations according to the direction of the association: forward-looking or backward-looking, respectively. The direction of the association can also be, in principle, determined for other that contiguous collocations, but this is decidedly a less straightforward procedure. For example, a stimulus-response pair *machine* → *translation* (Kaisa) matches a contiguous collocation *machine translation*, the direction of the association is XY: *machine* elicited *translation* and not the other way round. *Translation machine,* in contrast, is an unlikely combination though grammatically and semantically possible (not found in the BNC or COCA). A stimulus-response pair *separate* → *each other* ("separate from each other maybe", Kaisa) which matches not only *languages separated from each other* but also *languages separated from one another* can be assigned to an XY association category, but with a bit lower degree of confidence. The proportional relationship between XY and YX associations in Matching MWU S-responses will be discussed in Section 6.3.5.

Some of the stimulus-response pairs are not a verbatim match to the usage patterns of the corpus. However, a more approximate correspondence is identifiable. It seems possible to match a stimulus-response to the usage pattern by making either a semantic or a structural abstraction: that of a semantic preference or a colligation in Sinclair's terms. Let us first take some examples of a semantic preference.

Kaisa associates *establish* with *correspondence* commenting "also from the text". However, in practice she uses *establish* and *correspondence* together only once, as can be seen from the concordance lines for the lemma ESTABLISH in (6.21).

(6.21)

1 **The quality of the sound changes** can better be **established** on the basis of numerous typological,
2 areas of research: cognate identification and **establishing phylogenic relationships between languages**.
3 ified for the sake of demonstration. 2nd Step: **Establish sound correspondences** The second step in
4 intend to demonstrate with my work. In order to **establish** the aptness of using two-level rules for testing

(Kaisa, C1, concordances for the lemma ESTABLISH)

Yet, the concordance reveals two other related usages where the objects of ESTABLISH are noun phrases *the quality of the sound changes* and *phylogenic relationships between languages.* Both of the instances can be grouped under the common topic 'correspondence between languages'.

Hertta associates *sized* with *different* and elaborates on her association, saying: "probably because there are different sized people and probably [I] have used sized in somewhere in my text and now it feels very funny that I've done that because it looks very funny". In her drafts, *sized* is only used with *medium*, in *medium sized* (6.22).

(6.22)

1 ized ungulates (Meso ungulates), large and **medium sized** mammals (Mega and meso mammalian) and
2 such as large ungulates (Mega ungulate), **medium sized** ungulates (Meso ungulates), large and medium
3 elk bones. Sheep, goat and pig fall into **medium sized** ungulates. Minimum Number of Individuals is with

(Hertta, C1, concordances for the word *sized*)

In the BNC, the three most frequent collocates of *sized* are *medium*, *small* and *different.* Thus, the evidence taken together suggests that *sized* may have a semantic preference for adjectives (and adverbs) assessing the degree to which something is larger or smaller (other frequent collocates of *sized* in the BNC are: *moderately, reasonable, suitably, decent, generously, good).*

Linda produces *time* in response to *during,* explaining her association in the following way: "usually if you're during something it means time". *During* occurs 26 times in her drafts, 20 of which feature the pattern *during the * interviews*. The remaining 6 instances are provided in (6.23).

(6.23)

1 d in the Federal Press Office in Berlin, Germany **during July-August 2010**. I will finish my thesis by the
2 to conduct eight interviews personally in Berlin **during June-August 2010**. I decided to conduct semi-stru
3 te between public relations and public diplomacy. **During the research process**, the focus shifted between
4 ny. Confidentiality: All information collected **during the study period** will be kept strictly confidential
5 ce depends largely on the chancellor. For example **during the years** of Chancellor Kohl some of the tasks o
6 o handle the publicity of the Russian Federation **during their presidency of the G8**. [ref.] Also,

(Linda, C1, an extract from the concordances for the word *during*)

In all of these remaining instances of the word *during*, some kind of reference to time can be noticed, suggesting a semantic preference. One other association of a similarly abstract nature produced by Linda is *embassies → foreign* ("foreign embassies"): in her drafts she writes about *German* or *Germany's embassies.*

Maisa associates *registered* with *births.* Example (6.24) shows the concordances for the lemma REGISTER both as a noun and a verb retrieved from Maisa's corpus.

(6.24)

1 was registered. Also **orphanhood information** was **registered** concerning all members of the household.
2 mated that approximately **76% of all births** were **registered** in the system, thus it was necessary to use

3 **births**, **deaths and migration** among the parishioners were **registered**. [ref.] The parish register d
4 in 2001, also the **sex and age** of the person was registered. Also orphanhood information was registered
5 registered. [ref.] The **parish register** data are still available, and have been useful whe
6 [ref.] have been using the **parish registers** in their research on mortality in Ovamboland. Th
7 [ref.] The same problem with the **civil register system** applies also to the case of mortality. In t
8 they began to use the Scandinavian **parish register system** in them. This means that the births, death
9 demographers, was the founding of **parish registers**. When the Finnish missionaries founded congregate

(Maisa, C1, concordances for the word family with the headword REGISTER)

Apart from *births* in lines 2 and 3, also *orphanhood information, deaths, migration, sex and age of the person* are used as objects of the verb REGISTER. The lines featuring REGISTER as a noun reveal that it is *parish* or *civil register* which are in question. So the verb REGISTER in the sense of making a record in a *civil register* has a semantic preference for demographic information.

At the same time, *registered → births* as well as *establish → correspondence* could be categorised as collocations because these pairs of words do occur verbatim too. However, whether it should be deemed a limitation or not, a WAT presupposes a specific word as a response to a stimulus word, so even if a stimulus word has a semantic preference for a set of words in the respondent's mind, she is forced to choose one word from this set. In the same way some colligational responses, abstractions in the structural or grammatical dimension, can also be sometimes interpreted as both representing a verbatim association and a more general pattern.

For example, Hertta associates *seem* with *to be* ("something seems to be something"). This stimulus-response pair can be categorised both as a collocation and as a colligation: while SEEM + *to*-inf. is a structure frequently used in the respondent's drafts (48 instances), with different verbs in the role of the infinitive such as *form, cover, follow, prevent, represent*, *be* is the most frequent verb representing this structure (21 instances).

A similar case is demonstrated by Hertta's other association, *relatively – old* ("the stuff I am doing is relatively old"). *Relatively old* occurs in her drafts as an exact wording, but a more general structure is a colligation of *relatively* with an adjective (see 6.25).

(6.25)

1 in the material from 381 individuals. Women are **relatively** more **abundant** in ages 15- 20 years and men
2 an even higher age of 8 years. This would mean a **relatively old** animal [ref.]. PICTURE OF THE REPRES
3 on from pre-Christian to Christian rites has been **relatively rapid** in Finland and the Scandinavian

(Hertta, C1, concordances for the word *relatively*)

At the same time there are stimulus-response pairs which do not bear an exact correspondence with the usage patterns but can be explained with the notion of a colligation. For example, Kaisa associates *intend* with *to do* ("intend to do something"). However, in all three occurrences where INTEND colligates with the infinitive (6.26), she collocates INTEND with the verb DEMONSTRATE.

(6.26)

1	ical linguistics. On the whole, **I intend to demonstrate** the need for formal	kaisa_10_300410.txt
2	rical linguistics. On the whole **I intend to demonstrate** the need for formal	kaisa_4_100809.txt
3	, non-standard tasks as well, as **I intend to demonstrate** with my work. In o	kaisa_6_300909.txt
4	l implementation of the model was **intended** for analysis, the model is bidi	kaisa_6_300909.txt

(Kaisa, C1, concordances for the lemma INTEND)

In the same way, Hertta comes up with *to do* in response to *difficult* ("I don't know if I felt that this was difficult to do"). In her drafts, there is just one case where *difficult* is used in a different structure, in all the other instances *difficult* colligates with the infinitive in a predicative construction, as can be seen from the concordance lines in (6.27).

(6.27)

1 etation of a grave and its contents will be more **difficult** than, for example, to say that a man who is
2 Sweden since 1323 (chapter 3.3). It is, however**, difficult to know** if Catholic Lutheran or Orthodox
3 picture of what religious should be. The term is **difficult to define** because it concerns things that are inta
4 dug through the grave. In such cases it has been **difficult to determine** what finds belong to the filling
5 98). This overlapping of burial rituals makes it **difficult to distinguish** between the two inhumation burial t
6 bones (and other grave goods) to the graves are **difficult to establish**. Bird and fish bones are mentioned
7 n traditions in Christianity by saying that it is **difficult to lapse** of the old ways even if one has adapted
8 terial. For example the closeness to a Church is **difficult to prove** especially on the poorly studied rural ma
9 The occurrence of wooden constructions was **difficult to sort out** because of the lack of documentation

(Hertta, C1, concordances for the word *difficult*)

Here are two other examples which are quite clearly colligational. Kaisa produces the stimulus-response pair *concerning → this* ("also a phrase"). *Concerning* always occurs with a noun phrase in her corpus, as (6.28) shows.

(6.28)

1 tics of the data influence the decisions I took **concerning the architecture of the implementation**. A finite
2 al and archaeological sources give only hints **concerning the contact and relation between the Sumerians**
3 ovided, and since I will not get into details **concerning the etymology of the words**, I will make a compro

(Kaisa, C1, concordances for the word *concerning*)

Hertta in response to *indicates* writes: *something* ("so something indicates something"). In her writing samples, 10 times out of its 18 occurrences, the lemma INDICATE colligates with a noun phrase. These co-occurrences are presented in (6.29).

(6.29)

1 ioned animal bones in grave contexts. This could **indicate a lack of interest** in animal bones by the excavat
2 uity. The animal bones from the cemetery in Turku **indicate a late town burial** where the site was in
3 hamlet burials as late as the 15th century, could **indicate a more relaxed attitude** towards pre-Christian tr
4 s a criterion of a Christian burial, but it might **indicate Christian influence** [ref.]
5 onze Age-early Iron Age and a dwelling place has **indicated a longer interest** to the Luistari area [ref.]
6 ne decomposes quickly but traces of pottery could **indicate food offerings** for the dead. Animal sacrifices
7 were mostly over 12 to 18 months of age, 2 bones **indicating an individual** over 4,5 years. One bone was
8 goat. Also large ungulates and mammals had marks **indicating butchery**. The marks appear on vertebra,
9 ls under 5 years of age. The ages of the animals **indicate …normal distribution of animals** in … . It is poss
10 aves had a rectangular form and a flat base. This **indicates a coffin burial**. To the north of the cemetery,

(Hertta, C1, concordances for lemma INDICATE co-occurring with an NP)

Undeniably, in this group of Matching MWU S-responses, collocational associations by far outnumber those of semantic preference or colligation. It is possible that collocational associations or verbatim associations are stronger than more abstract associations, or, more specifically, strong enough to be elicited in a decontextualised task such as a WAT (see Section 6.3.3 for a more detailed discussion based on quantitative evidence). At the same time, it should be taken into account that it is methodologically much harder to identify the more abstract associations, and WAT in particular seems to favour verbatim associations. However that may be, the data discussed in this section shows that more abstract associations described by Sinclair as colligation and semantic preference are psycholinguistically real.

6.2.2. Non-matching MWU S-responses

Some of the responses which are categorised as syntagmatic do not match the usage patterns. However, either the respondent's comments or searches in general corpora indicate that the association is syntagmatic as it forms a MWU but a different one from C1 usage patterns.. Such responses very often quite forcibly remind us that the respondents have lives outside their Master's theses, where they also use English. What is especially striking about Non-matching MWU S-responses is that on the whole we would expect the responses to be biased towards the academic domain since all the stimulus words are taken from the respondents' writing samples. From this perspective, the Non-matching MWU S-responses manage to outweigh the more expected associations. Let us look at some of these associations in Table 6.2.

Table 6.2 Examples of Non-matching MWU S-responses from different data sets

Student	Stimulus-response	Respondent's comment and other notes	Corresponding MWU in C1	
			MWU	N of occurrences
Kaisa	*account → bank*		TAKE *into account*	5
	relationship → with		*relationship between*	7
			genetic relationship	11
	source → target	"has this relation like if you have a source then you have a target, like in translation" [works in a translation agency]	*open(-)source*	6
	bilingual → children		*bilingual data*	3
	paper → term	[apparently a term paper is what she has to write as part of her studies]	*paper-and-pencil linguistics*	16
	apt → student	"well that is a movie I have recently seen"	*apt for*	3
	sound → wave		*sound change(s)*	54
			sound correspondence(s)	29
	for → me		*for the sake of*	5
			for instance	9
Hertta	*time → table*		*at the same time*	4
			time periods	4
			long time	5
	mind → trick		KEEP/BEAR *in mind*	4
	building → site		*church building*	8
	fall → behind	"I don't know if I am feeling that I am falling behind because I have to work sometimes and can't really write when I want to"	FALL *into* (types)	
	missing → person	"missing person, I don't know maybe I watched too many CSA episodes"	BE *missing from*	
Linda	*set → mind*	"set or mind set, actually"	SET *out to find/examine*	6
			set/s of + plural noun	9
	task → hard		*task oriented*	5
	strong → light		*strong position*	4
			strong emphasis	2

Maisa	*access* → *Microsoft*	“software access” [Microsoft Access is a programme from Microsoft]	*access to*	5[90]
	distribution → *food*	“distribution of stuff or food”	*age distribution/HIV*	7
	transmission → *joy division*!	is a ... of joy division [is a song by Joy Division]	*mother-to-child transmission*	7

It seems clear that in some of the examples it is not the salience of the non-academic domain which overrides the thesis-related associations but something else. Let us take three examples from different students which look remarkably similar in this respect: *time* → *table* vs. *at the same time* (used 4 times) and *set* → *mind* vs. SET *out to find/examine* (used 6 times), *paper* → *term* vs. *paper-and-pencil linguistics* (used 16 times). In the first two examples, the more expected word combinations are also way more frequent in general English than the combinations which were in fact elicited: in the BNC *at the same time* is about 17 times more frequent than *timetable(s),* and SET *out* + inf. is more than 20 times more frequent that *mind set(s).*[91] However, intuitively it seems quite natural that *at the same time,* SET *out* + inf. or *paper-and-pencil linguistics* were not produced as associations to *time, set* and *paper.* The reason for this, formulated by Sinclair (1987), even though he did not conduct any experimental studies, is that it is the most independent sense of a word that comes to mind first. In the units *timetable*, *mind set* and *term paper,* the most independent senses of *time*, *set* and *paper* are retained better than in the units *at the same time*, *paper-and-pencil linguistics,* SET *out* + inf., where they are delexicalised to a degree that may start to hinder spontaneous syntagmatic association in the absence of context or intention to express a certain meaning. This decontextualisation is of course a characteristic feature of a WAT. In other words, in cases when we are presented with only one word from a unit of meaning whose semantic prosody is not given and this one word participates in the unit as heavily delexicalised, it is very hard to predict the rest of the words participating in the same unit, that is, the full form of the unit will not be available to intuition (see Section 2.6.4 for the theoretical discussion of the question). In line with Sinclair’s hypothesis, I will call this phenomenon the *core meaning effect.* The effect will be discussed further in Section 6.2.3, where it manifests itself most strongly. Section 6.3.1 will summarise all the evidence available on the effect and provide some quantitative data too.

[90] 5 out of 5 occurrences of *access.*

[91] All the possible spelling variants of the compounds *timetable* and *mind set*, as one word, two words, or hyphenated, are taken into account.

6.2.3. Non-matching MWU M-responses

Meaning-based associations elicited by the stimulus words which in fact combine with other words to form MWUs in the respondent's own written production are especially intriguing. In accordance with the hypothesis it should be easier from the processing point of view to give syntagmatic associations if they are available. Thus, these responses should have been syntagmatic, but for some reason they ended up being meaning-based. It is the reasons why the expected syntagmatic association was not elicited that this section will be concerned with.

Some of the 'non-conforming' Non-Matching MWU M-responses can be grouped under the name 'ignored prepositions'. In this group the responses are meaning-based while the respective stimulus words actually co-occur with certain prepositions in the respondents' texts. So it seems that respondents 'ignored' the prepositions which could have been supplied as syntagmatic responses and went on to interpret the meaning of the stimulus words. Sometimes it is the meaning of a stimulus word together with the preposition it occurs with which is interpreted as if the respondent took it for granted that the stimulus word went with the preposition. For example, in response to *sake* Kaisa produced *reason* ("gives a reason"), but of course *sake* means 'reason' in the unit *for the sake of.* In (6.30), there are more examples from the 'ignored prepositions' group.

(6.30)

Kaisa: *dealing → handling:* DEAL *with; definition → description: definition of; equivalents → equals: equivalents to; followed → consecutive: followed by; instead → opposition: instead of; mapped - linked: mapped to; similar → the same: similar to; purpose → reason: purpose of; lack –none: lack of.*

Linda: *consistent → sure: consistent with; summary → conclusion: in summary; findings → results: findings of.*

Maisa: *documented → archives: documented in; contribute → give*: CONTRIBUTE *to*; *affected - effect: affected by; majority → minority: majority of.*

The second group, presented in (6.31) is formed by responses which, in order to be syntagmatic, would have to violate the seemingly more natural forward-looking direction of the association. They would have to be YX associations.

(6.31)

Kaisa: *file → input: grammar file*; *method → way: comparative method*; *implementation → program: computational implementation*; *transducer → automaton: finite(-)state transducer(s).*

Linda: *material → stuff: research* or *interview material(s); skills → talent: language skills* (8 out of 8 occurrences of *skills*).

Hertta: *ungulates → animals: large* or *medium sized ungulates; constructions → buildings: wooden constructions (11).*

Maisa: *ratio → rate: dependency ratio; design → fashion: sample design.*

Taken together with the evidence from the Matching MWU S-responses where XY responses are dominant, a preference for a meaning-based response instead of a YX syntagmatic response strongly suggests that the natural tendency is to predict what will come next rather than what must have preceded a certain word. Such a conclusion would be in line with the idea of syntagmatic prospection (Sinclair and Mauranen 2006). In contrast, 'syntagmatic retrospection' does not seem to be intuitively plausible.

The third group of Non-matching M-responses which deserves separate attention is the *core meaning effect* group mentioned in the previous section. In this group, the fact that a stimulus word elicits a meaning-based response despite entering into a co-occurrence relationship in the respondent's own written production can be a result of the respondent's reaction to the core meaning of the word instead of its co-occurrence patterns due to a high degree of delexicalisation in those patterns. I will now describe three examples of the core meaning effect in detail and then list some further examples from different respondents as I have been doing with other groups of responses so far. The proportion of the Non-matching M-responses which can be explained by the core meaning effect in percentages will be presented in Section 6.3.1.

In my first example, in response to the stimulus word *number* Kaisa supplies an association *digit* which is a meaning-based response. However, in her own writing *number* is considerably often used in the pattern *a number of* (5 occurrences). *A number of* is a quantifier (e.g. Sinclair 1990) which is used with plural noun groups and allows an intervening adjective like *a large/limited/surprising/substantial number of*. In the BNC the construction *a number of* with an optional adjective in the middle has the frequency of 196.21 instances per million words, which is substantial. The *Oxford Dictionary of English* mentions this construction separately under the second sense of the noun *number, "*a quantity or amount", defining its meaning as "several". So, indeed, *number* in the sense of 'several' is not an independent lexical item but requires the presence of other elements of the construction: it collocates with the indefinite article *a,* the preposition *of,* and colligates with a plural noun group and optionally with an adjective. The adjective-element is both a colligation and a semantic preference because adjectives which fit the construction, forming a colligation, belong to a particular semantic set which can describe or evaluate quantity. The examples from this semantic set range from such common variants as *large, small, limited, finite, fair, significant, huge, restricted, sufficient, growing, high, sizeable, considerable, remarkable, vast, rising, varying, certain* to quite creative usages as *prodigious, conspicuous, manageable, convenient, modest, confusing, respectable* (the examples come from the BNC). All of these features are present in Kaisa's usage, as can be seen from (6.32).

(6.32)

1 ations in the original Wiki concept, TWiki adds **a number of** features that make it suitable for e.g. projec
2 ledge and allows for rapid comparison of **a large number of** languages. [ref.] developed a probabilistic
3 to several other languages, and there have been **a number of** prevailing theories that have first generally
4 non-Assyrologists have come up with **a growing number of** Sumerian etymologies and found more gramma
5 n languages is to systematically present **a large number of** words found in the languages' basic vocabulary

(Kaisa, C1, concordances for the pattern *a number of*)

Yet, *number* presented separately in a WAT does not bring the construction to her mind. This core meaning effect suggests that perhaps a construction or a unit of meaning *a number of* is not stored under the 'entry' for *number* in the mind but has a separate 'entry' of its own. Representations in the mind seem to be organised according to most independent meaning rather according to form (see also the argument in Section 2.6.4). As such, *number* is stored linked to its most independent sense "an arithmetical value, expressed by a word, symbol, or figure" (*Oxford Dictionary of English* 2010), and *a (+adj.) number of + plural noun* in its complete form is stored with the functional meaning 'to refer to a certain quantity of countable items usually meaning several'. The association between these two 'entries' is not presupposed and should not be taken for granted.

In the second example, Kaisa associates *following* with *after*. Undoubtedly, it is a meaning-based response interpreting the meaning of *following*. It is not clear though whether Kaisa takes it to be a preposition with the meaning "coming after or as a result of", an adjective with the meaning "next in time" like in "*the following day"* or transforms it into *the following* in her mind which is a noun meaning "what follows or comes next" (*Oxford Dictionary of English* 2010). At any rate, in her own writing *following* is used in two main patterns both of which serve the metadiscoursal or, more specifically, metatextual (Mauranen 1993) function of "referring forward" (Sinclair 1990: 395) in the text. One of them is a common pattern in academic writing which emerges in Kaisa's writing as more precisely specified to the point of being fixed: *In the following* [to refer forward] + *I will* [to state the function of an intention] + a discourse label (Sinclair 2004 [1982]), such as *provide an overview* or *summarize some key studies,* which specifies the author's intention in terms of the discourse act she is going to perform. Example (6.33) provides the concordances for the concgram FOLLOWING/I.

(6.33)

1 actual theories of language affinity. **In the following I will provide an overview** of the relationship
2 helps refine NLP applications. **In the following I will summarize some key studies** where knowledge
3 and language reconstruction. **In the following I will present a few examples** of older and more
4 such morphological alternations, and **in the following I will concentrate on describing** the first
5 the accuracy of the implementation. **In the following I will provide an overview** on the relationship of
6 as much as it does the implementation. **In the following I will summarize some key studies** where
7 Indo-European, Germanic and Altaic. **In the following I will explain the structure** of the entries with
8 Rule-Variables, Sets and Definitions. **In the following I will go through an example** grammar from
9 several **computational implementations**. **In the following, I will introduce a few of them**.
10 describe next. The HFST-TWOLC Grammar **In the following, I will go through** the two-level grammars
11 Formulating Two-level Rules **In the following, I will elaborate on** the two-level grammars for

(Kaisa, C1, concordances for the concgram FOLLOWING/I)

The concordances show that the whole pattern is quite fixed. *In the following I will* occurs in the exact form without any intervening words. The phrases which follow the fixed part of the pattern exhibit variability, but can clearly be grouped under a common category. The category of semantic preference is not fit for this role as the commonality between the phrases is not of a semantic nature. There seems to be a need for a new category to accommodate this example. In pinning down the nature of the co-occurrence, two ideas from previous research seem to be particularly useful. One of them is the concept of discourse reflexivity (Mauranen 1993) which narrows down metadiscourse to text about text specifically. Indeed, what we have here is not related to evaluation or stance: the author is simply prospecting ahead and saying what the reader should be expecting to find in the text in the immediate future. In doing that she is compelled to name the discourse act she is going to perform: *concentrate on* X, *explain* X, *elaborate on* X, *go through* X, *introduce* X, and/or the element of the text (or the discourse item) she is talking about: *an overview, a few examples*. Here what comes in very handy is Sinclair's concept of a *discourse label* as a feature of the interactive plane of discourse which emphasises the aspect of "continuous negotiation between participants" in a discourse (Sinclair 2004 [1981]: 52).

What we seem to have here in terms of a unit of meaning is a collocation of *in the following*, which itself can be analysed as a collocation, with *I will* and something like a 'text-organising preference' for discourse labelling. Possibly, a 'text-organising preference' is a variant of semantic preference on the interactive plane of discourse, though this is just a guess. What is important for us here is that the frequency and consistency of the pattern suggest that it forms a unit of meaning for the writer with its components bound by syntagmatic association. However, *following* does not elicit *in the following* or *I will* in a WAT from the writer. Instead she interprets the meaning which can be inferred from the word *following* alone, i.e. the one which it has without the contribution of other words. The conclusion is that the sequence *In the following I will* + a discourse label has evolved into a separate unit which does not necessarily have a strong association with the word *following* itself.

The second pattern where *following* participates in the same writer's texts (6.34) is simpler but constitutes a unit as well.

(6.34)

1 for HFST-TwolC with only a few modifications. **The following description** of the grammar file is written
2 nalysis could not be performed unambiguously. **The following example** from [ref.] demonstrates this.
3 ial terms and definitions used in the field. **The following list** is adopted from [ref.]
4 rallel so no intermediate levels are needed. **The following picture** demonstrates how the system works:
5 TwolC is aimed to substitute Xerox's TwolC. **The following section** will cover HFST-TwolC in more detail.
6 rules which I use in the implementation. **The following sections** are by no means exhaustive accounts on
7 mar is the Rules section. A rule is built on **the following template**, already used in TwolC [ref.]
(8 possible because the content is well categorized **following** from the use of forms in all topics. TWiki)

(Kaisa, C1, concordances for the remaining occurrences of *following*)

As can be seen from (6.34), *following* collocates with the article *the* and has a preference for a set of discourse labels naming certain parts of an academic text: *description, example, list, picture, section.* The unit serves the function of a cross-reference (Sinclair 2004 [1982]: 54): it introduces the coming textual element to the reader. And again, this unit does not come to the mind of the respondent as an association to the word *following*.

The last example comes from Hertta's data set. She associates *rest* with *here* commenting "if it sort of a hope that I could rest somewhere". It is clear that she reacts to the verb REST meaning "relax" rather than to the noun meaning "the remaining part of something" (*Oxford Dictionary of English* 2010). This example is different from the previous as the core meaning effect seems to make the response biased not to the most independent sense of a word, but to the homograph whose meaning is more independent and interpretable from the word alone. Example (6.35) provides concordances for all the occurrences of the root *rest*.

(6.35)

1 lian Isthmus. The burial custom differs from **the rest of Finland** because eastern Finns are thought to fo
2 left to the primary production places whereas **the rest of the fish** was slated or otherwise treated for t
3 ö region on the western coast of Finland. In **the rest of the coastal region** inhumation burials start to
4 ood would also be covered by the same soil as **the rest of the grave**. Some animal remains were found
5 being 100-150 cm in length and 30-60 cm wide. **The rest of the graves** were 50-100 cm wide. Most of the
6 from Porvoo is a late town burial compared to **the rest of the material**. Animal bones do occur but their
7 ch closer to the material from Turku than to **the rest of the sites**. The problem with the town materials
8 The shallowest grave was only 10-15 cm deep, **the rest** being 20 to 50 cm in depth. The grave depth was
9 erovingian period grave 269 and the pig **mandible rests** were found in the middle of the grave near artefa
10 contain animal bones. The bones might be offal or **rests of bone handling**. In the site of the church of t
11 individual use. They are thought to be the **last resting place** and as such they reflect the thoughts abo

(Hertta, C1, concordances for the root *rest*)

As can be seen from (6.35), only in line 11 one of the senses of the first homograph is evoked. But even there, *rest* is used as part of a larger unit of meaning *the*/possessive pronoun *last/final resting place*, rather than with its more independent sense ‘relax’. Thus, the fact that a syntagmatic association with *place* is not elicited in WAT is also the result of the core meaning effect. At the same time, since the form used as a stimulus was *rest* rather than *resting*, *place* would have been a less expected syntagmatic response. For the same reason, the respondent was unlikely to make a link to the ‘anatomical’ *rests* either. What could have been expected is the production of the unit of meaning *the rest of* + NP since *rest* is almost exclusively used in this pattern (lines 1-7). But it does not come to mind, presumably because of the core meaning effect.

Further examples of the core meaning effect are provided in (6.36).

(6.36)

Kaisa: *addition* → *subtraction*: *in addition to* (3); *around* → *surround*: *around B.C.* (3); *bears* → *forest*: BEAR *in mind* (2); *hand* → *leg*: *on the one/other hand* (8); *hard* → *easy*: *It* BE *hard/harder to*-inf. (7); *important* → *non-trivial*: *It* BE *important to*-inf. (8); *least* → *most*: *at least* (13); *means* → *way*: *This means that* (7; used in the sense ‘consequence’, ‘result’); *order* → *realization*: *in order to* (21); *possible* → *impossible*: *It* BE *possible to*-inf. (>10); *rather* → *instead*: *rather than.*

Maisa: *reference* → *citation*: *Census Reference Night* (2); *overall* → *cloth*: always occurs in *the overall* + NP (5), in the sense “taking everything into account” (*ODE*).

Linda: *basic* → *normal: Basic Law* (8 out of 13 occurrences of *basic*); *find* → *catch: find out*; *if - when: if so;*[92] *implicit* → *explicit: implicit code of conduct* (8 out of all 8 occurrences of implicit); *no* → *yes: there is no need* (7); *present* → *now*: PRESENT/INTERNATIONALLY;

[92] *If so* means “if that is the case” (*Oxford Dictionary of English* 2010).

previous - before: PREVIOUS/RESEARCH; *prior → first*: PRIOR/RESEARCH; *situation → place: situation analysis* (10); *term → concept: short-(long-, mid-)term*; *there → here: there is no need*.

Hertta: *latest → new* ("if something is new, it is oftenly latest")/*at the latest* (3).

6.2.4. No MWU S-responses

Now we are moving on to responses given to the stimulus words which do not participate in any observable patterns in C1. In other words, these stimulus words function as relatively independent in the respondents' usage, and therefore, they are not expected to elicit syntagmatic responses. This does not completely exclude the possibility of a No MWU S-response because, even if a stimulus word does not exhibit any patterning in C1, it may have a dependent use in one of its other senses. Indeed, No MWU S-responses occur in the data but very rarely, as can be seen from Table 6.3, making No MWU S-responses the smallest group of all.

Table 6.3 No MWU S-responses

Type of response	Kaisa	Hertta	Maisa	Linda	Nora	Total
No MWU S-responses	5	4	1	0	8	18

Just like Non-matching MWU S-responses, the majority of No MWU S-responses come from the respondents' non-academic spheres of life: *every → day* ("every day", Kaisa); *coverage → broad* (Kaisa, uses *coverage* as a term in her writing); *early → morning*, (Kaisa, examples of her usage: *early 90's, implementations, approaches, scholars*); *actually → love* ("love actually, cause I like the movie", Kaisa); *only → child* ("somebody's ... someone is an only child", Hertta).

Other responses categorised as No MWU S-responses illustrate the category less clearly. There is a possibility that they are constructed on the spot as potential completions of a syntagm which contains the word given as a stimulus, i.e. on the open-choice principle, rather than retrieved by syntagmatic association. For example, Kaisa generated a stimulus-response pair *relatively → bad* with an explanation "relatively bad, as a phrase". *Relatively bad* is a possible word combination but by no means a common or even a frequent one: it does not occur in the BNC at all and appears in the COCA only 3 times. *Relatively* is an adverb modifier of degree, so an adjective is its legitimate company, but this is perhaps not enough to claim that the response *bad* demonstrates a colligation. The usage of *relatively* does not seem to be constrained to only positive or negative adjectives like e.g. that of *utterly*, so *bad* is not an indication of a negative attitudinal preference either. At the same time we do not have data from Kaisa's interaction in English, so we cannot be sure that a grammatically possible but infrequent combination has not become her habitual turn of phrase. In addition there is always a possibility of a recency effect: if she heard or produced this word combination just before taking the test, the likelihood of it being reproduced in the test on the idiom principle, i.e. by retrieval rather than construction, would have been very high. Since Kaisa herself commented "as a phrase", the response was categorised as an S-response. A similar example comes from Hertta's profile: *apparent → size* with a comment "someone has or something has an apparent size or apparently a size that might be apparent". In her texts *apparent* is used as a predicative adjective in a quite free pattern otherwise.

Another interesting example is *presumably → not* given by Hertta. In her explanation of the response she said: "I don't know if it is my assumptions are presumably not right or something but that is just how it came out". I find this example interesting because *presumably* is a relatively independent item which is used with a metadiscoursal function as was demonstrated in Section 2.3. It can also be uttered alone as a short answer to a question. Almost the only word which can meaningfully saturate an already independent *presumably* is *not*. *Presumably not* can stand as it is. An example from the BNC (6.37) illustrates the point.

(6.37)

the work of the housewife? Is that alienated labour? **Presumably not**: but it would be a very bold man, a Karl

(BNC)

Presumably not can also be representing an instance of a colligation since degree, modal and focusing adverbs are in general attested to frequently occur as single-word responses in spoken language (Carter and McCarthy 2006: 459-460), and an addition of *not* is a natural way to construct a negative answer. In other words an adverb combined with *not* can form an independent unit, and the fact of occurrence of an adverb plus *not* structure inside punctuation marks alone is proof of that. In (6.38), there are some more BNC examples of adverbs which appear with *not* separated from the co-text with punctuation marks on both sides: *apparently, surely, maybe, perhaps, definitely, certainly, evidently.*[93]

(6.38)

work would have been possible if these movements had not existed**? Absolutely not.** I followed the ideas of

[93] The query "_PUN _AV0 not _PUN" returned 458 hits in all.

an obsession with Koi a need to go completely Japanese**? Perhaps not —** but the Japanese garden tradition,
the BBC will not necessarily show a game every Sunday**. Perhaps not:** Wimbledon, Open Golf and even cricket
one of their watches. Trying to get a freebie**? Surely not!** Anyway, he didn't. A cold Catalan beauty told him

(BNC)

Some of the adverbs occur in this pattern much more often than others, e.g. *probably not* constitutes close to 20% of all the occurrences. So it is possible to suggest that *presumably not* can also be an approximation of a more frequent pattern like *probably not.*

Overall, the infrequency of the responses in this category supports the hypothesis that words independent in use do not elicit syntagmatic responses. Some of the No MWU S-responses relate to the dependent use of a stimulus word in its different sense, and some may in fact be constructed on the open choice principle. Another observation is that WA responses indeed seem to have a strong tendency to be directed towards saturation of meaning. This is also in line with the previous suggestions (see e.g. Section 6.1.2).

6.2.5. No MWU M-responses

No MWU M-responses are as natural, according to the hypothesis, as Matching S-responses. If a word is used with a meaning relatively independent from its surroundings, it is consistent with the theory that this meaning will be salient for the WA response because first, it is interpretable, and second, it does not have a clear or dominant pattern to be supplied as a syntagmatic response.

One word type which often falls into this category is an adverb. This is not surprising: for example in Sinclair (1990), adverbs are subsumed under the category for adjuncts together with prepositional phrases, which indicates that they tend to function relatively independently. To illustrate this, I have taken all No MWU adverbs which received meaning-based responses from Linda’s and Kaisa’s WATs and categorised them according to the types of meanings they express in Table 6.4.

Table 6.4 (All) No MWU adverbs which received M-responses in Linda’s and Kaisa’s WATs

Adverbs	Type of meaning[94]	Linda	Kaisa
adjuncts	time, frequency & place	*abroad → internationally* *immediately → now*	*recently → yesterday* *already → now* *often → a lot* *seldom → often*
	manner	*briefly → shortly*	*directly → straight* *thoroughly → well*

[94] Adapted from Carter and McCarthy (2006: 456-458) and Sinclair (1990).

	degree	*somewhat → certain*	
	focusing	*especially → specially*	*especially → mostly* *merely - just* *generally –mostly* *mostly → often* *specifically → precise* *altogether → all in all*
Disjuncts and conjuncts	modal/ evaluative/ viewpoint	*probably → maybe* *obviously → clear*	*certainly → surely* *clearly -transparently* *unfortunately → luck*
	linking	*similarly → same* *consequently → as a result* *conversely → however* *hence → thus* *nevertheless → however*	*yet → still* *however → even though* *moreover → still* *thus -that's why* *also → plus*
	Other disjuncts		*traditionally → old* *theoretically → practically* *usually* (uses as a disjunct in the beginning of a sentence) → *often*

Table 6.4 shows that disjuncts and linking adverbs frequently elicit M-responses. Probably this happens because they are less integrated into the sentence and are therefore even more clearly independent, which is also emphasised by their being separated from the rest of the sentence by punctuation marks. On the whole, the main point the examples provided in the table are intended to show is that M-responses are given to stimulus words whose meaning is interpretable alone without the help of other words.

In this section, Section 6.2, WA responses have been compared to C1 usage patterns qualitatively. The observations made so far will be developed further and tested quantitatively in the next section.

6.3. Revisiting the main tendencies observed

Several effects and tendencies were noticed while categorising WA responses into Matching, Non-Matching and No-MWU in Section 6.2. In this section, I group the accumulating evidence according to the specific questions emerging from the categorisation, such as the core meaning effect and the relative strength of different types of association. I also provide quantitative data where possible. In Sections 6.3.8 and 6.3.9, the data from the comparison of usage patterns with the priming language is also brought in to examine the possible continuity between the three types of data.

6.3.1. Core meaning effect

In Section 6.2.3, the core meaning effect was used to explain Non-matching MWU M-responses. Basically the main hypothesis predicts that since units of meaning are psycholinguistically based on syntagmatic association, words which participate in units of meaning in use will elicit syntagmatic responses in a WAT. However, the core meaning effect shows that there is a competing force. When a word has a relatively independent sense and at the same time enters into a co-occurrence relationship as significantly delexicalised, our immediate impulse would be to react to the word's independent sense when it is presented alone, like in a WAT. This is a hypothesis which was formulated by Sinclair, but he did not have suitable data to show this tendency:

> The "core" meaning of a word – the one that first comes to mind for most people – will not normally be a delexical one. A likely hypothesis is that the "core" meaning is the most frequent independent sense. This hypothesis would have to be extensively tested, but if proved to hold good then it would help to explain the discrepancy [...] between the most frequent sense and what intuition suggests is the most important or central one. (Sinclair 1987: 323)

The strength of the core meaning effect seems to directly depend on the degree of delexicalisation. It is not by chance that in his later work, Sinclair calls a unit of meaning a meaning-shift unit. The fact is that whenever a word starts to participate in a co-occurrence relationship, it is always delexicalised in this relationship. But it is the degree of delexicalisation which would determine the predictability of the pattern. For example, in the Edinburgh Associative Thesaurus (EAT), a database of WA responses where each stimulus word was presented to about 100 native speakers of English, the most popular response (41 out of 100) to *light* is *dark* which is a meaning-based response reflecting the core meaning of the word.[95] Some of the less popular but still recurrent responses are syntagmatic, like: *bulb* (8), *bright* (5), *house* (4). However, in these word combinations *light* preserves its meaning to a large extent. In contrast, there is no response *in the light of* or *bring to light* or *shed light on* even among one-off responses. If the degree of delexicalisation did not play a significant role,

[95] Interestingly, in comparison to 41 people responding *dark*, only 7 people responded *heavy* which is an interpretation of the meaning in the sense of 'weight'. This is also a manifestation of the core meaning effect since *light* has a noun form which has a more independent meaning that *light* as an adjective in the sense of 'not being heavy'.

we would not get syntagmatic responses in a WAT at all because it is decontextualised and there is no access to the meaning of the target phrase: respondents provide their answers only on the basis of the form of the stimulus word and meaning which is contained in it.

In this study, the WATs are biased towards the respondents' texts as all the stimulus words are taken directly from them. The respondents are, as it were, put into a context which is likely to stimulate Matching MWU S-responses and these responses are expected. Non-matching MWU M-responses are not expected, so these responses provide good data to look for the reality of the core meaning effect. Table 6.5 shows the proportion of Non-matching MWU M-responses which can be explained by the core meaning effect.[96]

Table 6.5 Proportion of Non-matching MWU M-responses explained by the core meaning effect

Type of response	Kaisa WAT5	Kaisa WAT6	Kaisa, total	Hertta	Maisa	Linda	Nora	Total
Non-matching MWU M-responses	39	30	69	13	23	31	18	154
Non-matching MWU M-responses which can be explained by the core meaning effect	11	13	24	2	3	11	5	45
	28%	43%	35%	15%	13%	35%	28%	29%

Table 6.5 shows that all respondents gave WA responses where the core meaning effect could be identified, yet the proportions of them vary. Since this variability is displayed in Kaisa's WATs too, it is unlikely to be caused by the differences in processing between students. It is more likely that the stimulus words producing the core meaning effect were represented in different proportions in different WATs since they were selected from the students' text more or less at random (see Section 4.2.5 for stimuli selection procedures).

6.3.2. Does syntagmatic association develop only inside a unit of meaning?

As observed in Section 6.2.1, most of the Matching MWU S-responses are analysable with the components of a unit of meaning postulated by Sinclair. This fact strongly suggests that syntagmatic association takes place inside the boundaries of a unit of meaning. "Items that

[96] Similar arguments appear in usage-based theories. In presenting different types of linguistic evidence for chunking, Joan Bybee (2002) points out that "the morphemes or words inside a chunk become autonomous from other instances. For example, speakers probably do not associate *go* in *gonna* with the lexical movement verb anymore. Sosa and MacFarlane (2002) show that subjects have difficulty identifying the word *of* when it occurs in frequent chunks such as *sort of* or *kind of*" (112). This last examples sounds particularly in line with the observation made in this study that sometimes WAT respondents produce a meaning-based response as if reacting to the phrase in which a stimulus word participates rather than the stimulus word itself (e.g. *sake* vs. *for the sake of,* see Section 6.2.3).

are used together fuse together" (Bybee 2002: 112), but, it seems reasonable to hypothesise, in order to fuse together they need to obtain a larger meaning that would unite them into a phrase, and this unifying meaning together with the proximity would result in a syntagmatic association which facilitates spreading activation. In Section 2.8, this argument was applied to the lexical priming theory, according to which any linguistic element can become associated with any other linguistic element if they co-occur frequently enough. It was argued that *of* does not seem to be closely associated with *the,* although *of the* is the most frequent bigram in almost any corpus. In this section I will turn to those syntagmatic stimulus-response pairs which match instances of usage found in the respondent's corpus but cannot be explained in the unit of meaning paradigm. It is thought that even if a stimulus-response pair matches just one instance of usage, this correspondence cannot be dismissed as a coincidence. It seems that in order to accommodate the examples below, either the concept of a unit has to be stretched to include other possible components, or it has to be acknowledged that syntagmatic association is possible also outside the boundaries of a unit of meaning.

Bybee (2002) argues for the primacy of sequentiality joined with the frequency of occurrence over constituent structure in emergent chunking. In more general terms, semantic coherence in this account does not seem to be an indispensable condition for chunking to occur. Bybee states that it is those items that are used together which eventually end up being chunked rather than those which belong to the same constituent, like NP or VP. To illustrate her point, she gives some examples of contractions like *I'll* or *I'm,* where the fact of contraction itself shows that chunking can occur across the boundaries of constituents. The auxiliary which is an element of a VP contracts with the NP rather than with the following verb of the VP apparently just because the most frequent verb which can follow the auxiliary in the construction is not even half as frequent as *I* which precedes the auxiliary. Constituent structure is not considered in this study, but one of the arguments with which Bybee supports her point is of immense and direct interest for the arguments of this study too:

> Humans from 12 months old to adulthood can learn repeated sequences of meaningless syllables, as shown by Saffran et al. 1996; Gomez and Gerken 1999, 2000. Moreover, Gomez has recently shown that both babies and adults can learn sequences of two nonce words that are separated by a third 'word' chosen from a large class (Gomez 2001). *Thus meaning is not necessarily involved in learning sequences*, suggesting that the basis for constituent structure may be recurring sequences and not just semantics. (Bybee 2002: 124, emphasis mine)

As will be shown with the examples below, it indeed seems that there is nothing in our cognition that would stop 'meaningless' associations, i.e. associations hinged on mere proximity. Another matter is that words which do not mean anything as a unit but just happened to occur close to each other rarely reoccur in discourse often enough to become syntagmatically associated. This brings us to the idea that the reasons behind the tendency of syntagmatic association to happen inside the boundaries of a unit of meaning lie at the level of language in discourse rather than language in the mind (see the discussion of psycholinguistic vs. other 'realities' of linguistic entities in Sections 2.8.3 and 2.11): only those elements have a chance to occur together often enough to ensure psycholinguistic entrenchment which are able to make a meaningful contribution to the discourse. But let us turn to the examples themselves.

The first group of examples I would like to focus on has a very clear structure. This structure is reminiscent of a binomial, an expression consisting of two words of the same class linked with the conjunction *and* (Malkiel 1959; Mollin 2011). However, here the word combinations of this structural type are by no means fixed or idiomatic. It also requires a considerable stretch of imagination to think of a certain unifying meaning which would consolidate the two participating words in a phrase. Their co-occurrence looks more like a coincidence or a result of an open-choice decision. Yet, the fact that it is reproduced in a WAT cannot be ignored.

Linda's data set is particularly rich in this type of associations. For example, she associates *image* with *reputation* commenting "thesis", and indeed the combination *image and reputation* occurs in her texts ten times and in the reverse order – *reputation and image* – seven times. The association *validity* → *reliability* seems to have the same explanation behind it: *reliability and validity* has five instances in Linda's corpus. The association *ideology* → *issues* even gets a comment: "thesis, political issues or something there was" which in a way contradicts the corpus data supporting the idea that WA responses, especially syntagmatic ones, can rely on the implicit memory in comparison to the comments which are always based on the contents of the declarative memory. In the corpus, the two words co-occur in the combination *ideology and issue positions/category*.

Some responses do not lend themselves to such grouping as easily, even though they do appear in the structure X *and* Y. For example, one of Linda's stimulus-response pairs is *office* → *person.* Her comment clarifies the association: "two categories in one model, Chancellor as a person and as an office, office – person, I see the graph". At first sight, it appears to be a meaning-based response reflecting a type of a binary non-gradable antonymous relationship like the one between *man* and *woman.* At the same time the respondent clearly says: “I see the graph”, so the association can also be based on proximity, even though *office* and *person* do not form a MWU. Linda’s writing samples confirm than *office* and *person* indeed co-occur, though their co-occurrence is not phrase-like, as can be seen from (6.39).

(6.39)

1 [ref.] Furthermore, no particular **office** or a single **person** that is directly responsible
2 official status. The two aspects – **person** and **office** – are on opposite sides of a continuum, but the
3 are two sides to the Chancellor: a **person** and an **office**. As suggested in the second category,
4 can be portrayed as a **person** and as an **office**. In domestic communication, the emphasis is more
5 one that presents her as a **person** (The Press **Office** of the German government, 2011b) and one that

(Linda, C1, concordances for the concgram PERSON/OFFICE)

Other examples of a similar relationship are Linda’s stimulus-response pairs *qualitative* → *quantitative* and *operative* → *strategic* (“and then the opposite will kind of be strategic, well, not the opposite, but one step down”) and Kaisa’s *back* → *front.* All of them look like meaning-based associations, yet it has to be taken into account that they occur together as well, as can be seen from Examples (6.40), (6.41) and (6.42). In the case of *operative* → *strategic,* there is also an alternative association which Linda could have used as a basis for a syntagmatic association. *Operative planning* occurs eight times in the corpus.

(6.40)

1 [ref.] write that it is possible to combine **qualitative** and **quantitative** methods, for example use
2 [ref.] write, it is possible to combine **qualitative** and **quantitative** methods together. The
3 in which order the methods are used (first **qualitative** and then **quantitative**, vice versa or together)
4 reliability and validity are not the same in **qualitative** research as in **quantitative**. It is not likely to
5 researcher was a Finn. I decided to conduct **qualitative** research, even though **quantitative** research
6 that there is more freedom in the analysis of a **qualitative** research material than **quantitative** research
7 and **quantitative** methods, for example use **qualitative** methods to create variables for a quantitative

(Linda, C1, concordances for the concgram QUALITATIVE/QUANTITATIVE)

(6.41)

1 the research problem, one has to consider the **operative** and **strategic** communication planning for the
2 these concepts are used to discuss the **operative** and **strategic** communication with regards to a
3 The difference between **strategic** and **operative** communication planning lies within the questions:
4 in **strategic** planning, which is the base for **operative** planning. Then the next three sections focus more

(Linda, C1, concordances for the concgram OPERATIVE/STRATEGIC)

(6.42)

1 Sets are typically character classes like e.g. **back** vowels, **front** vowels, consonants or liquids, such
2 There are three types of vowels: **front**, **back** and neutral. Each **front** vowel [y, ö, ä] has a
3 change *k > h, the h, only appears in **front** of **back** vowels (a, o, u). In set4 *k remains k. Thus, it
4 or the backness of the word so that **front** and **back** vowels do not appear together in a single,
5 the vowel of the first syllable is a **front** or a **back** vowel. In practice, the ambiguity arising from
6 syllable determines the **frontness** or the **backness** of the word so that front and back vowels do

(Kaisa, C1, concordances for the concgram BACK/FRONT)

Some other allegedly syntagmatic responses seem to be even more abstract that those reflecting colligations or semantic preferences. They are reminiscent of Susan Hunston's idea of semantic sequences which are thought of as "sequences of meaning elements rather than ... formal sequences" (Hunston 2008: 271). These were not originally hypothesised as psycholinguistically relevant units but it seems from the WATs that it is not impossible to have a semantic sequence represented in the mind. An interesting example comes from Linda's data set. When cued with the stimulus word *comment*, she answered *say* and explained "and then you say something". In her corpus COMMENT strongly co-occurs with direct speech: out of all 15 instances of COMMENT, in 11 cases it is followed by direct speech as in Example (6.43).

(6.43)

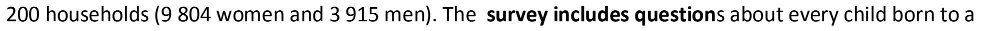
a task oriented, not as an entertainer. The next **comment** illustrates this: **"Her mind is that 'I do my work**

(Linda, C1)

In Hertta's data set, *supports → idea* ("somebody supports an idea so I wrote an idea") and *discovered → somewhere* ("something might be discovered somewhere") also look like examples of semantic sequences.

Some other Matching MWU S-responses which do not fit the model of a unit of meaning cannot be grouped in any meaningful way. These associations seem to be based on mere proximity with apparently the factor of recency coming to the fore. For example, in response to *survey* Maisa produced an association *questions* and commented on her response by saying "survey includes questions". *Survey questions* looks like a possible collocation and the phrase occurs in the BNC for instance, but it does not occur in Maisa's reference corpus as such. In contrast, in her own corpus of texts the most frequent collocation is *sentinel survey* (appears 14 times). But in the draft she wrote just before taking the test (which was at least two days before), the following line (6.44) can be found.

(6.44)


200 households (9 804 women and 3 915 men). The **survey includes question**s about every child born to a

(Hertta, C1)

As it can be seen, the comment repeats the line verbatim.

A similar thing happened with one of Linda's WA responses. She produced a stimulus-response pair *private → intimate* and then commented: "It was actually a sentence that I deleted yesterday like maybe half an hour before I had to have it in, I don't like that sentence, well, I had to delete something and deleted a sentence where a guy said that private information is also very intimate information I was like I don't really know what you mean so I deleted, so I guess it is stuck in my head, that sentence". The actual sentence she was talking about turned out to be a bit different, but different in a predictable way: *An image provides detailed and personal information, and is therefore very intimate. [Reference].* So it was *personal* rather than *private* which was used in the sentence, but in Linda's memory it became semantically approximated. The corresponding collocation which could have been fallen back on is *private life:* Linda used it 10 times in her thesis drafts as a contiguous XY collocation (or, in other words, as a concgram with the configuration AB), and therefore it could be a good candidate for syntagmatic association. Yet the recency factor seems to have overridden the more frequent and semantically coherent association.

One more example of a Matching MWU S-response which deserves attention is Maisa's *aids → HIV*. It could have been categorised as an M-response, but the sequence *HIV/AIDS* appears a substantial number of 44 times in her writing, therefore it may have developed a syntagmatic association by proximity.

In all, there are not too many WA responses which do not comfortably fit into the unit of meaning framework: most of them have been analysed or at least mentioned above. While these responses do not discredit the psycholinguistic reality of a unit of meaning, they suggest that proximity plays a very important if not decisive role for syntagmatic association. The importance of proximity is further supported by the fact that it is contiguous collocations which are reproduced in WATs most often. This tendency will be described in the following two sections.

6.3.3 Collocational response vs. semantic preference or colligation

As we have seen in Section 6.3.1, which described Matching MWU S-responses, all the components of a unit of meaning appear in WATs: syntagmatic association which has long been displayed in WA research can be not only verbatim but also of a more abstract kind. Syntagmatic association is abstracted or approximated either grammatically, resulting in a colligation or semantically resulting in a semantic preference. It was also noticed that a handful of seemingly syntagmatic responses cannot be analysed within the unit of meaning framework. The proportions of all these types of syntagmatic responses are shown in Table 6.6.

Table 6.6 Types of syntagmatic responses

Type of response	Kaisa WAT5	Kaisa WAT6	(Kaisa, total)	Hertta	Maisa	Linda	Nora	Total
Matching MWU S-responses:	16	50	66	57	25	32	39	219
	100%	100%	100%	100%	100%	100%	100%	100%
Collocation	14	43	57	19	17	21	29	143
	88%	86%	86%	33%	68%	66%	75%	65%
Colligation	1	3	4	24	1	0	6	35
	6%	6%	6%	42%	4%	0	15%	16%
Semantic preference	1	4	5	11	4	2	2	24
	6%	8%	8%	19%	16%	6%	5%	11%
Semantic sequence	0	0	0	2	0	1	2	5
	0	0	0	4%	0	3%	5%	2%
Other	0	0	0	1	3	8	0	12
	0	0	0	2%	12%	25%	0	6%

With the reservation that it is methodologically more difficult to identify more abstract syntagmatic associations, it is quite clear from Table 6.6 that collocation as a type of syntagmatic association by far outnumbers all the other types. The only exception from this trend is Hertta's profile in which colligations are even more popular than collocations. An interesting case is also exhibited in Linda's profile, whose "Other" associations, i.e. those which cannot be assigned to the familiar types of responses, take up a quarter of all the Matching MWU S-responses she has. The likely explanation of these two outlying preferences is the nature of a WAT: however carefully the stimulus words are arranged in a way to preclude any sequencing effects, respondents are prone to be influenced by the previous responses they give and get attracted by a certain strategy of answering the test items (see Section 4.1.5). For example, out of the 24 colligational responses Hertta gives, eight are of the type *someone, something, somebody*, e.g. *suggest → something.* At the same time it is also indisputable that at least in the analysed WAT Hertta's syntagmatic associations are of a more abstract kind that that of e.g. Kaisa. Anyway, collocation is a popular type of a syntagmatic response with all the respondents which suggests that verbatim associations are more likely to be reproduced in a WAT and therefore, presumably, they are stronger from the point of view of syntagmatic association than the more abstract ones.

6.3.4. Contiguity and the strength of representation

Just like it is collocations that are reproduced in WATs most often, it is contiguous collocations that are predominant, as can be seen from Table 6.7.

Table 6.7 Contiguous and non-contiguous collocations in collocational Matching MWU S-responses

Type of response	Kaisa WAT5	Kaisa WAT6	(Kaisa, total)	Hertta	Maisa	Linda	Nora	Total
Collocational Matching MWU S-responses:	14	43	57	19	17	21	29	143
	100%	100%	100%	100%	100%	100%	100%	100%
Contiguous	12	38	50	14	15	18	22	119
	86%	88%	88%	74%	88%	86%	76%	83%
Non-contiguous	2	5	7	5	2	3	7	24
	14%	12%	12%	26%	12%	14%	24%	17%

In other words, if a word is associated with a specific collocate (rather than a colligation or a semantic preference), it tends to be immediately to the left or to the right of it. If we evoke an already used comparison of a unit of meaning with the structure of an atom, we may say that the most ‘invariable’ electrons are closest to the core. In more general terms, as it was pointed out in Section 6.3.2, proximity or contiguity seems to be highly important for spontaneous association and perhaps for spreading activation. It is also possible that a collocation can form *because* of the proximity of two words in the first place as it was suggested in the aforementioned section.

6.3.5. The direction of syntagmatic association

As suggested in Section 6.2.1, contiguous collocation as a response type in a WAT can be either prospective or forward-looking (XY) or retrospective or backward-looking (YX). A prospective association where a stimulus word elicits a word which follows it seems to be intuitively more natural. Analysis of Matching MWU S-responses enables testing of this hypothesis. As can be seen from Table 6.8, all the respondents except Maisa have come up with more forward-looking associations than backward-looking, both if only contiguous collocations, for which it is easier to determine the direction of the association, are taken into account, and if all Matching MWU S-responses are counted irrespective of the type.

Table 6.8 The direction of association in contiguous collocations and in all Matching MWU S-responses

Type of response	Kaisa WAT5	Kaisa WAT6	(Kaisa, total)	Hertta	Maisa	Linda	Nora	Total
Contiguous collocations in Matching MWU S-responses:	12	38	50	14	15	18	22	119
	100%	100%	100%	100%	100%	100%	100%	100%
XY	7	34	41	12	7	12	21	93
	58%	89%	80%	86%	47%	67%	95%	78%
YX	5	4	9	2	8	6	1	26
	42%	11%	20%	14%	53%	33%	5%	22%
All Matching MWU S-responses:	16	50	66	57	25	32	39	219
	100%	100%	100%	100%	100%	100%	100%	100%
XY	10	41	51	41	10	18	35	155
	63%	82%	77%	72%	40%	56%	90%	71%
YX	5	4	9	11	12	9	2	43
	31%	8%	14%	19%	48%	28%	5%	19%
both	1	5	6	5	3	5	2	21
	6%	10%	9%	9%	12%	16%	5%	10%

In order to see what actually happens when the direction of association comes into play, I will look at Maisa's profile, which is slightly non-conforming to the general tendency, and compare her XY and YX associations. Maisa has 15 contiguous collocations among her Matching MWU S-responses. Out of these 15 responses, seven are forward-looking and eight are backward-looking. Table 6.9 lists all these stimulus-response pairs along with their list numbers, which show the order in which stimuli appear in the WAT, and the number of times the corresponding collocations occur in the corpus. The list number is important because, as has been mentioned before, preceding responses can influence the following ones, and the number of occurrences can shed light on the strength of a collocation.

Table 6.9 XY and YX collocations in Maisa's data set

XY			YX		
Stimulus-response pair	List N	N of C1 inst.	Stimulus-response pair	List N	N of C1 inst.
overnight → travellers	11	1	*epidemic → HIV*	3	13
census → data	15	8	*errors → sampling*	6	2
spectrum → software	36	3	*instance → for*	14	5
sentinel → survey	45	27	*projection → population*	19	19
derived → from	60	1	*structure → population*	48	11
methods → section	70	1	*migration → international*	59	4
crude → rate	84	1	*infection → HIV*	90	7
			prevalence → HIV	103	63

Let us first look at YX associations. Even though some of the responses recur: *population* is a response to two stimuli, and *HIV* is a response to three stimuli; the list numbers of the pairs do not suggest any obvious influences. It is also possible to say that interference of the preceding responses is a less likely basis for the eventual responses than syntagmatic collocational association because almost all the stimulus-response pairs occur quite frequently as contiguous collocations in Maisa's writing samples, which suggests that their strength of association for her must be quite salient. Thus, it is possible that the syntagmatic associations were made in spite of the unnatural backward direction due to the high strength of association.

In contrast, many of the collocations corresponding to the XY associations occur just once in Maisa's texts.[97] The recency factor could have played a role here, but it is also quite plausible that the possibility of making a forward-looking association facilitates giving a syntagmatic response, and, therefore, a syntagmatic response becomes possible even if the association is relatively weak.

The data of special interest is presented in the lower part of the table. Four out of seven XY collocations shown in Table 6.9: *census → data; spectrum → software; sentinel → survey* and *methods → section* happened to be tested in the reverse order as well, that is, Y-collocates *data, software, survey* and *section* were presented as stimulus-words. These stimuli together with the responses they received are presented in (6.45).

(6.45)

survey → questions (N96, Matching MWU S-response, other)
software → computers (N51, Non-matching MWU M-response)
section → a part (N100, Non-matching MWU M-response)
data → figures (N27, excluded due to the possible priming effect: N4 *demographic → figures*)

As can be seen, in none of these four cases did Maisa produce a collocational response, presumably because of the backward-looking direction of association she would have to make. Even in the case of a relatively strong collocation in her C1 *sentinel survey*, *survey* did not elicit *sentinel.* It would be possible to suggest that the order of presentation could have played a role in case Y-collocates *data, software, survey* and *section* were presented first and

[97] They were still classified as Matching MWU S-responses because as it was argued earlier the fact that a spontaneous WA response matches even one occurrence in the texts can hardly be regarded as a coincidence.

activated a priming effect when X-collocates *census, spectrum, sentinel* and *methods* were presented next, causing the production of collocational associations . But this did not happen. In fact, collocates *census, spectrum, sentinel* and *methods* were presented first and elicited collocational responses, and then Y-collocates *data, software, survey* and *section* were presented but did not elicit collocational responses *in spite of* the priming effect. For example, *spectrum* elicited *software* as an XY association at number 36 in the test, but further down the test, at number 51, *software* failed to elicit *spectrum* as an YX association, but was responded with a Non-matching M-response *computers* instead.

Further evidence of the facilitatory effect of forward-looking associations and the difficulty of making a backward-association comes from analysing Non-matching MWU M-responses, i.e. those responses which could and, according to the hypothesis, should have been syntagmatic but did not live up to the expectations. Some of them, as was discussed in Section 6.2.3, can be explained by the unnaturalness of associating backward.

In all, the material analysed strongly suggests that the possibility of making a forward-looking or prospective association has a psycholinguistically facilitatory effect for the elicitation of a syntagmatic response in a WAT. In contrast, the necessity of making a backward or retrospective association, as it were, has a somewhat inhibitory effect: syntagmatic associations were not elicited in many cases where they could have been expected. On the basis of this data, it seems possible to claim that prospection is more natural for syntagmatic association than retrospection.

6.3.6. Statistical significance of the connection between WA responses and C1

The analysis has shown that different components of a unit of meaning can be elicited in a WAT, which speaks in favour of the psycholinguistic reality of the model. In other words, linguistic elements that co-occur and comprise a unit of meaning seem to be psychologically associated as well. However, how systematic is this connection? In order to give an answer to this question, it is necessary to reformulate it.

Linguistic elements which function as components of a unit of meaning are dependent on each other for the meaning they can only express together. So when a word participates in a unit of meaning in the student's writing, it can be considered dependent on the accompanying co-text in usage and incomplete in terms of meaning. When a word does not apparently participate in any units of meaning in the student's writing, it can be considered relatively independent in usage. In the same way, when a word elicits a syntagmatic response in a WAT, it can be considered dependent in the respondent's mind. If it elicits a meaning-based response, it has a relatively independent meaning of its own for the respondent. It follows that our question can be reformulated as: Do the words which reveal themselves as dependent in usage also tend to be dependent in the respondent's mind? That is, do the words which participate in units of meaning in usage also tend to elicit syntagmatic associations in a WAT? And the other way round, do the words which are used relatively independently also tend to reveal themselves as independent in the respondent's mind? That is, do the words which do not participate in any apparent units of meaning in usage also tend to elicit meaning-based associations?

The categories used in the comparison of usage patterns with WA responses can be expressed as follows (see Table 6.10). The number of Matching MWU S-responses is the number of words which showed themselves as dependent both in usage and in a WAT. Since the category of Non-matching MWU S-responses implies that a word was actually used in a unit of meaning in the respondent's texts, but in a WAT a different syntagmatic association was elicited which also forms a unit of meaning that is attested in the language, even though not in the respondent's text (like, *bilingual children* instead of *bilingual data*). These responses are also counted as dependent words both in usage and the WAT, just like Matching MWU S-responses. Therefore, in Table 6.10 a separate row shows the total number of words which are dependent both in usage and the WAT (Matching MWU S-responses plus Non-matching MWU S-responses). The number of Non-matching MWU M-responses is the number of words which are dependent in usage but independent in the WAT. The number of No MWU S-responses is the number of words which are independent in usage but dependent in the WAT. And No MWU M-responses give us the number of words which are independent both in usage and the WAT. Thus, a simple 2x2 contingency table (see Figure 6.1) can be used to test the significance of the interrelationship.

	Dependent in WAT	Independent in WAT
Dependent in use		
Independent in use		

Figure 6.1 Contingency table used to test the interrelationship between WA responses and C1

Table 6.10 gives the numbers of associations in each of the groups. The last row shows the p values of the significance of the interrelationship calculated by using the Fisher's exact test.

Table 6.10 Responses by category and the significance of the connection between usage and WAT

Type of response	Kaisa WAT5	Kaisa WAT6	Kaisa, total	Hertta	Maisa	Linda	Nora	Total
Number of responses[98]	110	117	227	112	67[99]	99	102	607
Matching MWU S-responses	16	50	66	57	25	32	39	219
Non-matching MWU S-responses	12	24	36	23	4	12	13	88
Dependent in use and WAT (Matching MWU S-responses + Non-matching MWU S-responses)	28	74	102	80	29	44	52	307
Non-matching MWU M-responses (Dependent in use, independent in WAT)	39	30	69	13	23	31	18	154
No MWU S-responses (Independent in corpus, dependent in WAT)	2	3	5	4	1	0	8	18
No MWU M-responses (Independent in use and WAT)	41	10	51	15	14	24	24	128
two-tailed P value (in Fisher's exact test)	<0.0001	=0.0011	<0.0001	<0.0001	=0.0008	<0.0001	<0.0001	<0.0001

The interrelationship proves to be significant or even highly significant in all of the cases. That is, it is syntagmatic association which underlies co-occurrences observed in use and it is the same syntagmatic association which is elicited in WATs. This finding has the following implications. First, WAT provides data which is relevant for language use. Second, multi-word units the students used in their writing are indeed likely to be produced on the idiom principle as they are syntagmatically associated in the mind too, as shown by the WATs. The fact that when Kaisa's results from two WATs are collapsed, the p value becomes more significant rather than less significant suggests that if more WA responses were analysed, they would reveal an even stronger (rather than weaker) interrelationship. The variation in different students' proportions of responses in each category is likely to be due to uneven distribution of words with different properties in the students' WATs, as mentioned in Section 6.3.1 and explained in Section 4.1.5.

[98] Each student has a different number of responses analysed because one WAT, which was taken as a basis for analysis, could contain a slightly different number of stimulus words from 135 to 116 (in total 735 stimulus-response pairs were analysed). In all the data sets some stimulus-response pairs had to be excluded because of the suspected influence of the previous responses or insufficiency of data for making a classification decision: a stimulus was matched to just one occurrence in the corpus data.

[99] Quite many stimulus words had to be excluded because the respondent did not give any answer.

6.3.7. Is it implicit memory which is tapped?

In Chapter 2, components of a unit of meaning were hypothesised to be glued together by syntagmatic association. It was also claimed that it is this syntagmatic association which enables them to be produced as a single choice i.e. on the idiom principle. Arguably, syntagmatic association is something which is stored in implicit memory. This is assumed to be the reason why the pattern is not readily available to intuition or retrospection (Section 2.6.4). At the same time the fact that this 'information' is part of implicit memory rather than explicit accounts for the automaticity of operation on the idiom principle.

This study cannot give a direct answer to the question whether it is implicit memory which is involved in processing of units of meaning. The WATs were not timed, so there is no information on which responses took longer to arrive at and which were more spontaneous. It is not obvious, though, that such information could give a conclusive answer. However, this study provides some indirect evidence on the question. Implicit memory is supposed to be functioning spontaneously without loading conscious attention. Therefore, it can be assumed that if in retrospective comments the respondent states that a certain stimulus was hard to respond to, it suggests that the she had to go through certain conscious thinking process to arrive at the response. This is why I will look at stimulus-response pairs which were pronounced hard or difficult by the respondents themselves and see whether these responses tend to be meaning-based or syntagmatic, making allowance for the caveats related to retrospection, though.

All the stimulus-response pairs which were described as hard by the respondents in one way or another are presented in Table 6.11 (the actual comments provided by the students are supplied in Appendix C).

Table 6.11 Stimulus-response pairs which were 'hard' for the respondents

Student	Stimulus-response	M/S	Compared to C1	MWU(s) in C1
Maisa	*rural → countryside*	M	Non-matching	*rural populations* *rural and urban (areas)*
Nora	*important → for me*	S	Matching, colligation	*important for somebody/something*
	lays → somewhere	S	Matching, colligation	*(the) main focus* LAY *on* + NP (4 out of 5 occurrences of LAY)
	accepted → o.k.	M	No MWU	
	flexibility → not fixed	M	No MWU	
	underlies → lies under	M	No MWU	
	hereby → with this	M	No MWU	
Linda	*consistent → sure*	M	Non-matching	*consistent with*
	situation → place	M	Non-matching, core meaning	*situation analysis*

	there → here	M	Non-matching, core meaning	*There is no need to* + inf.
	somewhat → certain	M	No MWU	
	whereas → when	M	No MWU	
Hertta	*context - burial*	S	Matching, YX	*burial contexts*
	trial → excavation	S	Matching	*trial excavation(s)*
	adults → male	M	No MWU	
	child → infant	M	No MWU	
	vary → difference	M	No MWU	
	possibly → possibility	M	No MWU	
	located → on	S	Matching, colligation	BE *located at/behind/in/on*
Kaisa, WAT5	*so → that way*	M	Non-matching	*so far, so that*
WAT5	*otherwise → or*	M	Non-matching	a colligation with a negative
WAT5	*pinpoint → point out*	M	Non-matching	colligation with an object and a semantic preference for some kind of 'infelicities'
WAT5	*beyond →behind*	M	Non-matching	GO/SCOPE/EXPAND *way/far beyond* NP
WAT5	*would → should*	M	Non-matching	*would have*, WOULD/IF
WAT6	*along → by*	M	Non-matching	*the lines of*, *along with* + NP meaning 'together with'
WAT6	*have → hold*	M	Non-matching	*have been, have undergone, would have*
WAT5	*and → or*	M	No MWU	
WAT5, N103	*moreover → still*	M	No MWU	
WAT5, N112	*however → even though*	M	No MWU	
WAT5	*thus → that's why*	M	No MWU	
WAT5	*yet → still*	M	No MWU	
WAT5	*whereas → but*	M	No MWU	
WAT6	*whereas → but*	M	No MWU	
WAT6	*also → plus*	M	No MWU	
WAT6	*since → because*	M	No MWU	
WAT6	*specifically → precisely*	M	No MWU	

As can be seen from Table 6.11, the 'hard' responses strongly tend to be meaning-based. All of Maisa's, Linda's and Kaisa's and more than half of Nora's and Hertta's 'hard' responses are meaning-based. Only two Nora's responses and three Hertta's responses which are syntagmatic remain unaccounted for. It is interesting that out of these five syntagmatic responses, three are classified as colligations. This fact seems to be able to at least partly explain why they could be difficult: even if one has a syntagmatic association with a structural feature, it may be difficult to put this structural feature into words in a decontextualised WAT. Two stimulus-response pairs are left and both come from Hertta. One is *context → burial* and she said it was "surprisingly difficult because I talk about context all the time." The concordance lines in (6.46) reveal that the collocation is actually *burial context(s),* i.e. it was a YX or a backward-looking association which was already hypothesised to have an inhibiting effect.

(6.46)

1 sheep, goat or pig bones are found in **burial context**s the bones are subjected to the theory of ritual

2 may also be coincidental or accidental. **Burial context**s may also be disturbed by unrelated cultures,

3 that can be extracted from the **burial context**s. In the material used for this study there are

(Hertta, C1)

The last 'hard' stimulus-response pair which requires explanation is *trial → excavation.* It is a Matching, collocational, syntagmatic association. However, a closer look at her comment suggests that the collocation might still be part of Hertta's declarative knowledge. She says that earlier she used the phrase *test excavation,* but then this phrase was corrected by her supervisor, and now she knows that the correct variant is *trial excavation.* On the basis of this comment it is possible to infer that the collocation is not yet automatised and therefore did not come immediately to mind.

On the basis of this evidence it is possible to conclude that giving syntagmatic responses does not require extra effort while giving meaning-based responses can be quite demanding. This is far from being conclusive evidence that syntagmatic responses are prompted by implicit memory, but at least it suggests that this is a reasonable hypothesis.

6.3.8. Continuity between C1, C2 and WA responses

The analysis in this chapter showed that L2 usage patterns can be reproduced in WATs suggesting their holistic representation in the mind of the language users. In the previous chapter, it was shown that such usage patterns tend to come directly from the priming language suggesting holistic usage-based learning from exposure. Therefore, it is interesting to see whether there can be found any examples of C1 patterns for which both evidence of the source of acquisition and evidence of psycholinguistic representation have been tracked down through the two comparisons undertaken. Such examples could serve as an illustration of the continuity between the patterns of the priming language, usage patterns and WA responses for which the argument is made in this study using the findings of each comparison separately.

There are certain reservations which have to be spelled out before the data from the two comparisons can be brought together. First, not all words participating in significant concgrams of C1 are tested in WATs. Second, WA responses cannot be expected to always match the usage patterns for different reasons discussed in this chapter. Therefore any examples of the continuity between C1, C2 and WAs that can be presented are necessarily going to be selective. Yet the matching between the three types of data cannot be explained by pure chance.

Table 6.12 presents some examples of such usage patterns matching across the three types of data. Since the qualitative comparison of usage patterns with the priming language patterns was carried out only for two students, the table focuses on the patterns of these two students, Kaisa and Maisa. For example, the first line shows that the word *sentinel* co-occurs with the lemma SURVEY both in C1 (27 times) and in C2 (37 times) representing a collocation *sentinel survey(s)*, and when Maisa was presented with the word *sentinel* in her WAT, she responded with the word *survey*.

Table 6.12 Examples of continuity between exposure data, production data and WA responses

Co-occurring word	Co-occurring word	C1	C2	Pattern represented by the concgram	WAT	Student
SENTINEL	SURVEY(S)	27	37	*sentinel survey(s)*	*sentinel → survey*	Maisa
PROXIMATE	DETERMINANTS	19	2	*proximate determinants (of fertility)*	*determinants → proximate*	Maisa
GROWTH	RATE	18	8	*(population) growth rate*	*rate → growth*	Maisa
LIFE	EXPECTANCY	27	19	*life expectancy*	*expectancy → life*	Maisa
HIV/AIDS	IMPACT	15	7	*the impact of HIV/AIDS*	*impact → HIV*	Maisa
SIMILARITY	MEASURES	10	36	*(orthographic) similarity measures*	*similarity - measure*	Kaisa
COGNATE	PAIRS	15	33	*cognate pairs*	*cognate → pair*	Kaisa
FINITE	STATE	8	37	*finite-state (transducers, automata)*	*finite → state*	Kaisa
TRAINING	DATA	5	19	*the training data*	*training → data*	Kaisa
CLOSELY	RELATED	11	9	*closely related*	*closely → related*	Kaisa
AUTOMATIC	COGNATE	9	1	*automatic cognate (recognition,*	*automatic → recognition*	Kaisa

				detection, identification)		
COMPARATIVE	METHOD	12	8	*comparative method*	*comparative → method*	Kaisa
ACCOUNT	TAKEN	11	6	*taken into account*	*take → into account*	Maisa
EVEN	THOUGH	19	4	*even though*	*even → though*	Maisa
FOR	INSTANCE	5	6	*for instance*	*instance → for*	Maisa
DEPENDING	ON	5	16	*depending on*	*depends → on something*	Maisa
FOR	REASON	6	3	*for any/this/that/some reason, reason for*	*reason → for something*	Maisa
ASSUMPTIONS	ABOUT	13	27	*assumptions about* (e.g. *HIV, fertility*)	*assumptions → about HIV*	Maisa
INTO	ACCOUNT	5	9	TAKE *into account*	*take → into account*	Kaisa
CAUSED	BY	5	3	*caused by*	*caused → by*	Kaisa
CONCENTRATE	ON	6	2	CONCENTRATE *on*	*concentrate → on*	Kaisa

The table demonstrates remarkable continuity between the three types of data collected in the study: exposure data, usage data and psycholinguistic data. Since these three types of data are mutually supportive, cumulatively they make an even stronger point in favour of the availability of the idiom principle to second language users. The students in this study were able to acquire, use and store units of meaning on the idiom principle i.e. by syntagmatic association.

6.3.9. Are approximation and fixing psycholinguistically real?

Section 6.2.1 showed that syntagmatic association can be not only verbatim but also abstracted semantically or grammatically, forming the basis for semantic preference and colligation attested in usage. The same example can serve as evidence that the process of approximation in a unit of meaning suggested in Sections 3.5 and 3.6 and discussed in Section 5.5.1 is also psycholinguistically real. Here I would like to give one example which shows the psycholinguistic reality.

SPLIT/INTO (6/2) is an example of a concgram which matches C2 usage from Kaisa's data set. This is not surprising because the preposition *into* is a normal environment for the verb SPLIT: for example, it is the most significant preposition co-occurring with SPLIT in the BNC. However, when looking at the concordance lines of SPLIT*/INTO generated from C2 (6.47) and C1 (6.48), one notices an important difference:

(6.47)

1 Cognates False-Friends and Unrelated), and then **split** the first class **into** Cognates and False-
2 be accomplished by adding memory (M) or by **split**ting the input **into** smaller chunks (e.g.,
3 previous work on cognate identification can be **split into** three general areas of research:

(Kaisa, C2, concordances for the concgram SPLIT*/INTO)

(6.48)

1 example, rule 1 above could alternatively be **split into 2** separate rules: "Y realised as i
2 because each rule with variables needs to be **split into** as many subrules as needed in order
3 the way of decomposing the mapping was **split into two**: computational linguists used the
4 Automatic cognate recognition can be roughly **split into two** main areas of research:
5 family tree demonstrates that Proto-Finno-Ugric **split into two** subgroups: Finno-Permic and Ugric.
6 the way of decomposing the mapping was **split into two:** Computational linguists used the
7 approach to the comparison, I ended up **splitting** the implementation **into two** separate

(Kaisa, C1, concordances for the concgram SPLIT*/INTO)

While in C2 something is split into parts, in broad terms, in C1 it is *into two* 5 times out of 7 occurrences of SPLIT. This can be regarded as a case of fixing: the number of parts into which something is split becomes specific, or in other words a semantic preference for several parts becomes a collocation of SPLIT with *two*. *Two* is indeed a frequent collocate of SPLIT on the whole, it is in fact its fourth most significant collocate in the BNC. Yet, it is still remarkable that this specific feature revealed in usage is also reproduced in Kaisa's WAT: when she was prompted with the stimulus word *splitting*, she responded with an association "*two",* the fact which indicates that an association with *two* rather than a more abstract idea of 'several parts' has become represented in her mind too.

So, both approximation and fixing can be observed in WA responses too, which indicates that they reflect the processes that representations of lexico-grammatical patterns undergo in the mind.

6.4. Conclusions

This chapter compared WAs produced by the students with the usage patterns they exhibited in their C1s. It showed that the relationship between the response a word elicits in a WAT and the usage pattern it participates in is statistically significant. Words which are dependent in use tend to elicit syntagmatic responses, words which are independent in use tend to elicit meaning-based responses. That is, words which participate in units of meaning in use, initiate completion of the pattern in a WAT; words which function relatively independently have a meaning on which a WA response is then based. Therefore, the conclusion was made that it is possible to use WA responses as data which can inform us about the psycholinguistic processes working behind the corpus linguistically attested co-occurrences.

WAs produced by the students show that multi-word units they use in their writing are psycholinguistically supported by syntagmatic associations represented in the mind. That is, these multi-word units are produced by retrieval from memory on syntagmatic association, which is what operation on the idiom principle probably is. In other words, the idiom principle seems to work on syntagmatic association. The data analysis carried out in this chapter allowed to explore the properties of this syntagmatic association further.

In decontextualised tasks, like a WAT or any task where one needs to explain a word without any context provided, syntagmatic association can be hindered by the core meaning effect. The core meaning of a word is its most independent sense. It is this sense which gets interpreted in a meaning-based WA instead of a pattern completion in a syntagmatic response. It is difficult for a language user to provide a completion of a pattern in response to a single word in case this word participates in the pattern as significantly delexicalised, i.e. has lost most of its core meaning. This is why a pattern in question is called a meaning-shift unit.

Also, syntagmatic association seems to work largely inside the boundaries of a unit of meaning. However, there is evidence to suggest that it can be caused by mere proximity. That is, syntagmatic association can appear between items which co-occur but do not (yet) form a unit of meaning. It is an open question whether syntagmatic association caused by proximity can activate the development of a new unit of meaning or whether a unit of meaning has to form first to make this syntagmatic association entrenched. A certain number of WA responses pointing towards semantic sequences rather than units of meaning suggests that psycholinguistic predisposition to specific choices at higher textual levels is in principle also possible. But overall, it is not clear whether syntagmatic association has to be connected with a meaning to be represented in the mind or not. This is an interesting question which opens up avenues for future research.

Further, syntagmatic association seems to be strongest between collocates, i.e. words or other linguistic elements which co-occur verbatim. However, it is very important that a more abstract syntagmatic association, semantic preference or colligation, seem to be psycholinguistically real too. In this sense, the model of a unit of meaning postulated by Sinclair is psycholinguistically sound. The fact that meaning can be stored in more abstract representations than a word also points to variability and approximation inside a unit as possible and natural results of operation on the idiom principle.

The strength of syntagmatic association seems to be further influenced by (1) proximity: association between contiguous collocates is stronger than that between non-contiguous and (2) the direction of the association: the prospective or forward-looking association is facilitatory while ‘retrospective’ or backward-looking direction is in a way inhibitory. It also seems reasonable to suggest that the idiom principle makes use of implicit memory. However, this hypothesis, though theoretically sound, needs further testing.

The last two sections have brought together the evidence discussed in Chapter 5 and the present chapter, showing that there is continuity between the patterns of the priming language, the patterns used by the students, and the WA responses that therefore operation on the idiom principle occurs in acquisition, use and processing.

7. Conclusions

This study set out to explore the mechanism of the idiom principle in second language acquisition and use. With this aim in mind, the theoretical underpinnings of the idiom principle were analysed and its connection with second language acquisition and use revisited. New conceptual and methodological approaches to studying the operation on the idiom principle were developed. I will now briefly summarise the scope and the main arguments of each chapter before moving on to the bigger picture which emerges from this study.

Chapter 2 discussed Sinclair's conceptualisation of lexis and meaning with its major concepts of the idiom principle and a unit of meaning as the theoretical framework of this study. It raised the questions of dependent and independent uses of lexical items, core meaning, delexicalisation and meaning-shift. It was argued that the main differences of Sinclair's unit of meaning from other conceptualisations of a multi-word unit are: (1) its status of a new lexical item rather than a mere realisation of combinatorial possibilities of the words comprising it; (2) its incorporation of both syntagmatic and paradigmatic dimensions of choice and its ensuing tolerance of variability; (3) its ability to reconcile single and multi-word units. It was also suggested that the concepts of Sinclair's account of meaning should be regarded as interconnected components of one system. Thus, a unit of meaning was defined as an independent lexical item produced on the idiom principle, and semantic prosody as the communicative function of this unit of meaning. A link was made between language use and psycholinguistic processing with the suggestion that the idiom principle is a psycholinguistic mechanism of language processing which works implicitly by syntagmatic association and results in production or comprehension of units of meaning.

Chapter 3 reviewed the mainstream research on L2 acquisition and use of multi-word units. It demonstrated that the focus in this research is often on the errors L2 users typically make and the ways this situation can be remedied. The root of the problem is perceived to be in L2 learners' insensitivity to holistic processing, i.e. unavailability of the idiom principle or lack of it. It was suggested that, as Mauranen (2012) puts it, it might be more important to note that after all L2 learners get the multi-word units approximately right rather than slightly wrong (144). Since the changes they introduce into multi-word units are regular and can be explained cognitively by frequency effects and the superiority of memory for meaning over memory for form (see Sections 3.5 and 3.6), we might postulate approximation as a process which is inherent to the mechanism of the idiom principle.

Chapter 4 explained that, to test the hypothesis of availability of the idiom principle to L2 users, three types of data were collected and set against each other: individual corpora of Master's thesis drafts representing L2 usage patterns, word association responses representing representations in the mind and individual corpora of reference texts representing the priming language.

Chapter 5 worked with the comparison of L2 usage patterns to the priming language. It showed that 56% to 75% of significant patterns extracted from L2 usage corpora match the patterning of the priming language. The closer analysis of matching and non-matching patterns confirmed the reality of the approximation process and revealed a new process termed *fixing*.

Chapter 6 turned to the comparison of L2 usage patterns with word association responses. It was able to ascertain the relationship between the behaviour of a word in usage patterns and in word association responses, show the psycholinguistic reality of a model of a unit of meaning, and explore the properties of syntagmatic association further, examining the role of collocational association, contiguity, direction of association and the core meaning effect in its behaviour.

Together, the theoretical suggestions of Chapters 2 to 3 and the empirical observations of Chapters 5 to 6 lead to conclusions with respect to three aspects of the idiom principle: (1) L2 acquisition and use, (2) the model of a unit of meaning and (3) the processes behind the phraseological tendency of language. In what follows, I will look at each of these three aspects separately, drawing the findings together into a more comprehensive picture.

7.1. The availability of the idiom principle to second language users

In this study, it was shown that (1) more than half of significant L2 usage patterns also occur in their priming language; (2) the patterns match not only the patterning which is common for English in general but also field-specific uses; (3) the matching of the patterns can be very precise down to smallest detail; (4) the processes which bring about non-matching patterns, approximation and fixing, are intrinsic to the idiom principle; (5) L2 users exhibit extended unit of meaning patterning in their production; (6) this extended unit of meaning patterning is also reproduced in WATs demonstrating that units of meaning attested in use are also represented in the mind holistically; and (7) there is evidence for the continuity between the priming language patterns, the usage patterns and the word association responses. Taken together this is seen as evidence that operation on the idiom principle occurs in L2 acquisition, use and processing.

In other words, the phraseological ability of L2 speakers does not seem to be fundamentally different from NS ability. This study demonstrates that the idiom principle seems to be available to L2 speakers to a larger degree than is commonly assumed: in particular, L2 speakers seem to be able to acquire extended units of meaning from exposure, use them in context and hold them in memory on the idiom principle.

The processes of approximation and fixing detected in the data are not likely to be confined to L2 users only. The model of a unit of meaning predicts that these processes are normal for language use overall (see the discussion below). Yet the specific multi-word units produced by L2 users as a result of these processes may be typical of L2 use in particular. What can make their use of phraseological units different at times from NS use is the level of entrenchment of these units due to differently distributed exposure. The fact that these language users speak English as their L2 means that they have at least one more language in their repertoire. Therefore, their L2 serves them for only some of the functions a language usually serves: for example, for communication at work but not at home or vice versa, for handling some particular responsibilities at work but not all of them, the constellation of the usual contexts in which L2 is used can be very different. In short, the lexical patterns of L2 users may be different from Standard English even though they are produced on the idiom principle. For this reason, it may be useful to draw a distinction between being idiomatic in NS terms and operating on the idiom principle.

7.2. Developing the model of a unit of meaning

To bring together all the observations that emerged from the data with respect to the model of a unit of meaning and describe them cumulatively, I will use an analogy with an atom. It has already been evoked several times in passing (see Sections 2.2, 2.3, 6.3.4), but here it will be developed in more detail. Before I start, I would like to point out that in no way am I trying to suggest that the system of a language must resemble structures of the physical world. Then, why might drawing an analogy be helpful? First, it is a convenient and easily understandable, judging by my colleagues' comments, method of description which helps to bring all the features of a model together and clarify many complicated arguments. Second, we probably know much more about an atom than we do about a unit of meaning. Thus, the comparison raises useful questions which would not come to mind otherwise. For example, an atom can be positively or negatively charged, how about a unit of meaning then? Is it characterised by a similar feature? Importantly, the question does not have to be responded with a ‘yes’, it can just as well be: ‘no’: the analogy works as a way of generating questions.

Figure 7.1 below presents a schematic illustration of the analogy.

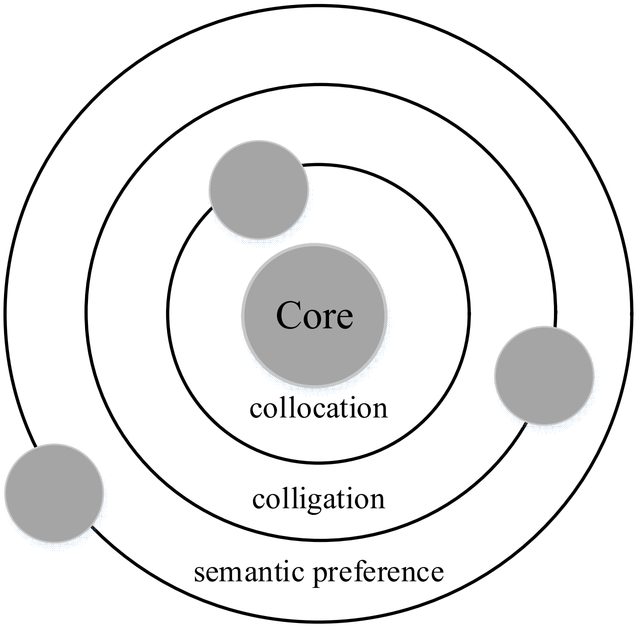


Figure 7.1 The model of a unit of meaning

The core is the invariable formal component, the one by which a unit is recognised. It resembles the nucleus at the centre of an atom. Semantic prosody is not marked in the figure, but it is the meaning which keeps the unit together. In terms of the natural sciences, the properties of the element can be determined in a chemical reaction. In the same way, extending the analogy, semantic prosody of the unit reveals itself when a unit of meaning is “put to use in a viable communication” (Sinclair 2004: 34). All the rest of the components of a unit of meaning are optional. In the figure, they are depicted as electrons in the electron cloud, each on its own energy level or orbital: that of collocation, colligation or semantic preference, which were categories in Sinclair’s terms. The figure gives an example of one ‘electron’ on each energy level. The fact that an orbital is not a planetary type of orbit along which an electron moves around the core, but the probability function indicating where we can find this electron suits the linguistic purposes well. There is no way to predict exactly how a component of a unit of meaning is going to be used.

One difference between units of meaning and atoms of the physical world is that while a unit of meaning consisting of the core only is a structural possibility, to the best of my knowledge there are not too many examples of atoms without electrons (except for the Hydrogen ion, perhaps). At the same time it is difficult to say how common units of meaning consisting of the core only are in language. In Section 2.3, it was argued that single-word units of meaning are possible. Yet, the core is not necessarily a word: it is defined as an invariable formal component, but the research demonstrates the possibility of variation even within one word. For example, Mauranen (2012) shows that approximation can occur inside single words without any impact on the meaning communicated: in the case of the produced item *successing* instead of *succeeding*, there is a structural approximation, and in the case of *negated* instead of *denied*, a semantic approximation can be observed, yet the communicative intention remains recognisable (101-102). So, the formal component by which we can identify a unit of meaning can be smaller than a word, and even items like *successing/succeeding* can be represented as the core plus associations.

To give another example, the proposed single-word unit of meaning *presumably,* discussed in Section 2.3, can also be regarded as an internally complex structure since it consists of a verb *presume*, which can be traced to Latin *prae* 'before' + *sumere* 'take' (*Oxford Dictionary of English* 2010) and suffix *–ly*. The question then is: what is it that forms the core? Since a unit of meaning is a lexical item produced on the idiom principle,[100] its formal realisation is basically determined by the speaker's/writer's cognitive representations. The core can be represented be any formal feature that for this language user reliably associates with the corresponding unit of meaning. The language user in question may approximate all possible components of this unit, but it will still be a unit if it is produced on the idiom principle, i.e. intended as a unit of meaning.

Let us now move on to the optional components. As the comparison of usage patterns with word associations showed, collocational, or verbatim, association seems to be stronger than that of semantic preference and colligation. Association also seems to be strongest in the case of a contiguous collocation, that is, when the collocational component is located

[100] It can also be said that it is a lexical item understood on the idiom principle by the hearer. Yet, it is difficult to perceive something as a unit of meaning if it was not first produced as a unit of meaning. Therefore, production is given the pride of place. Something produced on the idiom principle is not necessarily understood on the idiom principle: the hearer may try to work out the meaning of a multi-word unit by chopping it up into constituent components if the MWU in question is not familiar.

immediately before or after the core. Therefore, in the figure collocational orbital is placed closest to the core to reflect the strength of association. This does not contradict the structure of an atom. In an atom, the closer the electron is to the core, the stronger is the bond and the more difficult it is for the atom to lose this electron, which seems to be valid in the case of units of meaning too. In terms of a unit of meaning, the closer the component is to the core, the more fixed it is too. This postulate is in line with an estimation of a window size in which it is reasonable to look for collocates in corpora. Stubbs (1995) for example points out that spans 2:2 and 3:3 are often used and, according to Sinclair (1991), there is little collocational interest outside the span of 4:4. That is, collocations occur close to the core only.

Colligation, an association with a grammatical class, and semantic preference, an association with a semantic set, are then further away from the core than collocation. Indeed both of these categories involve a lot of positional and constituency variation. There is no evidence to date as to which of these components has a stronger association with the core. A reasonable hypothesis is that there are units of meaning with a strong semantic preference, and units of meaning with a strong colligation. But so far this is not reflected in the model. For the time being, they are given different orbitals because they are qualitatively different components, but the ordering of the orbitals does not mean that colligation is necessarily stronger than semantic preference. It is also important not to forget that these are optional components, therefore, not all units of meaning are going to have both or any of them, just like elements in the periodic table have a different number of orbitals.

The comparison of usage patterns with the priming language has revealed two processes: approximation and fixing. Through approximation, a verbatim association becomes loosened and moves to the category of semantic preference or colligation. Through fixing, in reverse, a specific alternative from a semantic set or a grammatical class becomes preferred, and what was a semantic preference or a colligation becomes a collocation. In terms of the model, this means that associations, just like electrons, can jump from one orbital to another, getting closer to the core or further away from it. With respect to the previous question of the location of semantic preference and colligation and their tentative placement on different orbitals, it is interesting to contemplate the possibility of semantic preference becoming a colligation and colligation becoming a semantic preference. The process of fixing also predicts that a unit of meaning can acquire new components. This means that new orbitals can appear (in case a unit did not first have components in the categories of semantic preference or colligation).

Another interesting point of comparison is that in an atom there is a particular order in which electrons fill the orbitals. In particular, electrons cannot appear on the second orbital, until the first one, i.e. the lowest, is filled up. For a unit of meaning, this would mean colligations and semantic preferences can develop only after a verbatim association is formed. This might suggest a meaningful hypothesis since it seems to be true that where there is a more abstract association, there tends to be a more easily identifiable verbatim association too. In other words, the fuzzier features often form a cloud which surrounds a more verbatim association. This can be postulated as a possible tendency.

7.3. The processes behind the phraseological tendency of language

According to the theoretical account presented in Chapter 2, when two or more words start to co-occur and associate with a specific communicative function, they are soon treated on the idiom principle as a holistic unit of meaning, an independent lexical item. Once the idiom principle switches on and replaces the open-choice principle, meaning-shift is launched: co-occurring words start to lose their core meanings, i.e. delexicalise, and become united by a shared semantic prosody gluing them together in a unit. The present study has found evidence for two more processes which seem to play an important role in the phraseological tendency of language: fixing and approximation.

Fixing, or the tendency of a pattern to become the preferred wording for an individual, as shown in this study, or for a language community, as it is reasonable to assume, is able to explain the accumulation of the instances of co-occurrence and the switch to being treated on the idiom principle. Figure 7.2 models the sequencing of stages through which a combination of words moves on its way to becoming a meaning-shift unit.

In the model, co-occurrence leads to fixing, which, in turn, leads to delexicalisation. Meaning-shift follows delexicalisation, but in fact it is a parallel process: a word which is delexicalising is shifting its meaning because it does not seem to be possible to lose one meaning without acquiring a new one, even if as a part of larger unit.

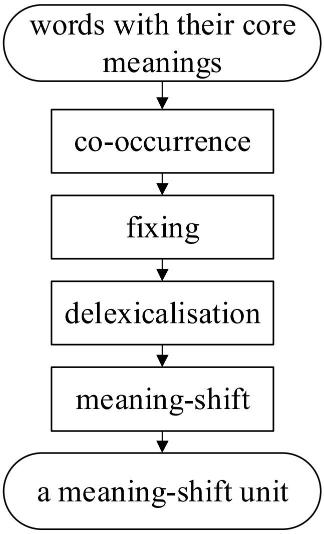


Figure 7.2 Delexicalisation and meaning shift

As proposed in Section 2.4, we can posit a continuum of delexicalisation and meaning shift. The delexicalisation of co-occurring words which have become a meaning-shift unit does not stop at that. The more these words occur as a unit, the more fixed their co-occurrence relationships become, and the larger is the meaning-shift possibly leading to a fixed and semantically non-transparent idiom. It is also likely that in the process of fixing the unit may attract new components, i.e. extend. Figure 7.3 is an attempt to visualise this continuum: it shows what happens after co-occurring words become a meaning-shift unit (MSU), i.e. it picks up where Figure 7.2 stopped.

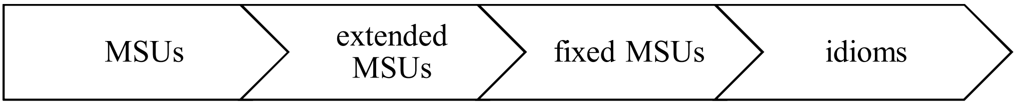


Figure 7.3 A continuum of delexicalisation

The present study has been diachronic only to a certain degree: its diachronic dimension lies in the fact that priming occurred before language use, and word association were collected both after the L2 users received certain priming and after they produced at least some of their written pieces. In other words, it is diachronic in terms of a change a lexical item undergoes in the process of acquisition by L2 users from the language they are exposed to, its subsequent usage and representation in the mind. Also, the drafts comprising C1 have been produced and collected longitudinally over a period of around a year.[101] It was possible to observe the diachronic in their nature processes of fixing and approximation when priming data was compared to usage data. However, I would not want to suggest that the pattern going through the process of fixing in individual usage only can also continue moving along the continuum of delexicalisation, that is, I do not believe that a fixing pattern can gradually become opaque in meaning without the approval of a discourse community. In other words, I would hypothesise that a unit of meaning can undergo a larger meaning shift only if it is picked up by others: this way the discourse community agrees on the 'legitimacy' of the new unit and acknowledges its new meaning. Therefore, in order to be able to track down further movement of a unit of meaning on the continuum of delexicalisation up to becoming an idiomatic expression, a different kind of data is needed: diachronic data of language use at a the communal level, rather than individual usage. In this sense decisions about meaning are made in the discourse (see Teubert 2005, 2010).

Going back to Figure 7.4, it should be said that delexicalisation can also be instantly reversed at any point along the continuum: the words comprising a MSU can be relexicalised. In this case, the meaning-shift unit loses its larger meaning, dissolves into constituent words again, and the words reclaim their core meanings.

The term *relexicalisation* is not new. For example, it is one of the key concepts in Partington (2006),[102] where he discusses it in the context of word play and laughter. To achieve a humorous effect, a joke-maker forces the hearer to reinterpret on the open-choice principle something which was first set up be read on the idiom principle, i.e. to relexicalise the delexicalised components of a multi-word unit.[103] Thus, Partington (2006) defines relexicalisation as "the 'freeing up' of the parts of a normally frozen, preconstructed lexical unit" (119). This is done by putting a multi-word unit or a part of it into a new co-text, or, in Hoey's terms, by overriding the habitual lexical priming.[104]

So Partington's examples illustrate deliberately evoked relexicalisation. But unintentional relexicalisation, in the sense of not seeking any humorous effect, also seems to be possible. If my interpretation is correct, the process Pitzl (2009, 2012) observed in English

[101] Yet, it must be mentioned that variation in lexico-grammatical patterning between different drafts was not observed, presumably because in the context of writing a particular text like a Master's thesis usage habits form and become fixed very quickly.

[102] See also Philip 2011.

[103] For example:
A: What happens if the parachute doesn't open?
B: That's known as 'jumping to a conclusion'. (Partington 2006: 119)

[104] Cf. Sinclair (1987): "Lexical choices which are unexpected in their environment will presumably occasion a switch" (324).

as a lingua franca use and called re-metaphorisation is a phenomenon similar to relexicalisation, or perhaps a direct consequence of relexicalisation in multi-word units whose meaning is based on a metaphor. She sees it as a process through which ELF users creatively "re-introduce metaphoricity" into idiomatic expressions by breaking them down into constituents and reviving the metaphorical images lying behind them (Pitzl 2009: 313, 316). Here are some of the examples Pitzl (2009) provides: *we should not wake up any dogs* (vs. *Let sleeping dogs lie*); *how to draw the limits* (vs. *to draw the line*); *put my hands into the fire for it* (vs. *de hand voor iemand in het vuur steken* in Dutch, the speaker's L1).

At the same time, we have seen that formal variation is not necessarily a sign of operation on the open-choice principle. While I would agree that when a speaker alters some formal features of a unit of meaning, this may trigger processes of relexicalisation and re-metaphorisation for the hearer, I would argue that the altered unit of meaning may still have been produced on the idiom principle. Semantic or structural resemblance to the original idiom (or any other unit of meaning) whether in English, or in any other language in the speaker's repertoire, is evidence that it was retrieved whole and intended as a unit even though it was approximated on the way. It is the communicative function of the approximated items like those in Pitzl's examples that indicate the production on the idiom rather than the open-choice principle. Since the communicative function stays the same as the original idioms had, and Pitzl shows this quite convincingly, the produced MWUs are instances of the same idioms, even though approximated. The meaning-shift has not occurred. Relexicalisation and re-metaphorisation would have required a change of meaning.

Thus, we can postulate two routes the process can take, and while one results in the decomposition of the meaning-shift unit into constituents, the other does not: the pattern loosens but does not dissolve and is still the product of operation on the idiom principle. Figure 7.4 illustrates the difference.

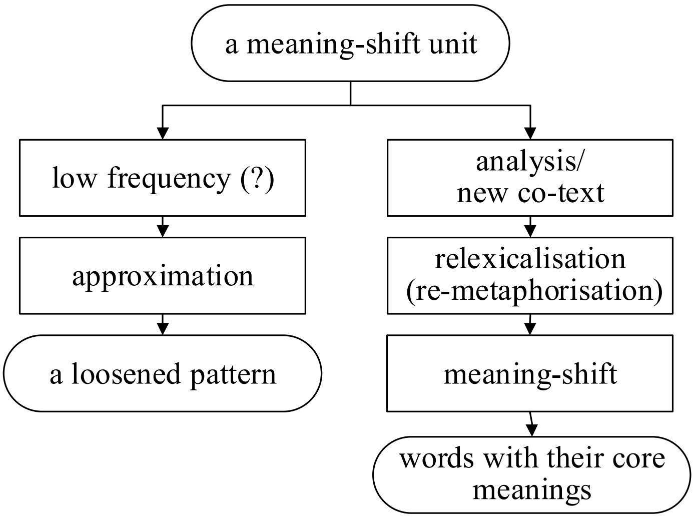


Figure 7.4 Approximation vs. relexicalisation (re-metaphorisation)

In other words, while relexicalisation presupposes a switch from the idiom principle to open-choice, approximation does not. And where there is no switch, there is no meaning-shift.

In all, as illustrated in the three figures of this section, meaning-shift can occur in two directions: towards continuous delexicalisation and in the opposite direction towards relexicalisation. The process of fixing assists delexicalisation and the formation of a meaning-shift unit. In contrast, the process of approximation does not necessarily lead to relexicalisation, at least on the part of the producer, since the overall meaning of the unit is retained.

7.4. Evaluation of the study

Perhaps the main contribution to linguistic research this study offers lies in its combination of three types of data, none of which has ever been used before in quite the same way: language use data, priming language data and a psycholinguistic type of data.

First, language use data was collected from five individuals over a relatively long period of time (a year, on average). At the same time, it belongs to the same genre, a Master's thesis. There are several advantages in this type of data: (1) it is naturally occurring, (2) it permits a developmental perspective, (3) its organisation into individual corpora of language production facilitates a cognitive approach.

Second, this study aimed at getting an insight into L2 users' acquisition and production of lexical patterns, at the same time adhering to usage-based explanations.

However, a usage-based perspective requires data about language exposure, but, as Hoey points out in his book on lexical priming, “a corpus […] represents no one’s experience of the language” (2005: 14), and therefore we are usually reduced to dealing with reference corpora which can only “indicate the kinds of data a language user might encounter in the course of being primed” (2005: 14). But how well can the BNC or the Bank of English or the COCA represent the kind of language that an L2 user is exposed to? This is a problem I tried to overcome by compiling individual reference corpora for the participants of this study which would be better representative of their particular primings.

Third, recently there has been an upsurge of interest in the psycholinguistic reality of corpus-based patterns, but in order to test this psycholinguistic reality, corpus data must be complemented with psycholinguistic data. This study demonstrated a possible approach to this task by developing the theoretically grounded methodology of comparing usage patterns and word association responses.

But together with the coveted advantages, this design brings several shortcomings too. The study is necessarily small-scale and includes data from five participants only. The collected corpora, being individual, are inevitably very small. These shortcomings together with the imperfection of some technical solutions may be responsible for the following problems.

The study may be biased towards verbatim co-occurrences at the expense of more abstract associations such as semantic preference and colligation. First, the operationalisation of units of meaning through concgrams and n-grams does not allow one to retrieve units of meaning based on more abstract associations only, without any verbatim co-occurrence of two words.[105] While it seems possible to hypothesise that formation of a verbatim association is a necessary stage in the development of a unit and thus, semantic preferences and colligations rarely exist without verbatim co-occurrence of certain elements, this hypothesis cannot be studied with the tools used in the present study. Second, the retrieval of units of meaning produced on the idiom principle is based on recurrence. Therefore, all the units of meaning which were produced just once remained behind the scenes. It is highly likely that these one-off units could exhibit more instances of abstract associations. Analysis of the data suggests that in individual language use, especially in a certain recurrent context, the process

[105] A possible alternative to the methodology used in this study would be to POS-tag the corpora which may have facilitated identification of colligations. This would have required asking rather different questions, though.

of fixing moves extremely fast: semantic preference or colligation may turn into collocations already from the second occurrence.

Also, the comparison of usage patterns and word association responses showed that contiguous collocations were reproduced in WATs much more often than non-contiguous. However, while ConcGram is designed to extract non-contiguous collocations, their number was quite small and the list of n-grams, contiguous co-occurrences, was almost just as useful as a list of concgrams. So it is possible that either the tools did not capture all non-contiguous co-occurrences, or the corpora were too small to exhibit a large number of them. On the basis of the present analysis, it was concluded that contiguous collocations are stronger than non-contiguous, and collocations are stronger than more abstract associations. This seems to fit in well with the model of a unit of meaning and the process of fixing postulated to be accompanying delexicalisation and meaning shift. It is also possible that a collocation forms because of the proximity of the two words in the first place.

7.5. The way forward

One of the main conclusions of this study is that the idiom principle is available to L2 users in acquisition, use and psycholinguistic representation of lexis to a much larger extent than is usually claimed. While in this study, fixing and approximation were the two processes which were able to account for much of the mismatch between the priming language and L2 production unexplainable by content and genre specificity, it is probably approximation which leads the researchers to assume production on the open-choice rather than the idiom principle. Both empirical observations and theoretical arguments presented here demonstrated that the process of approximation and the mechanism of the idiom principle are not mutually exclusive. On the contrary, approximation, just like fixing, is a natural part of operation on the idiom principle. At least this holds for language production.

The question which arises from these considerations is whether approximation triggers a switch to the open-choice principle on the part of the hearer/reader. Earlier studies have shown that variation introduced into multi-word units in ELF contexts does not cause any communicative disturbance (see Carey 2013; Mauranen 2005, 2012; Pitzl 2009, 2012; Seidlhofer 2009). However, we do not know whether the interlocutors had to switch to the open-choice principle in order to interpret the meaning of the units or whether they were able to process them on the idiom principle. It seems possible to hypothesise that there may be a difference between native speakers and L2 speakers when it comes to understanding approximated multi-word units: while the overriding of lexical priming, i.e. the certain novelty of accompanying forms, may cause NSs to switch to interpreting an approximated multi-word unit on the open-choice principle, just like in cases of word play (Partington 2006; Hoey 2005, especially, the 'drinking problem' hypothesis), L2 users may not notice the novelty, due to lower levels of entrenchment, and continue processing on the idiom principle. In this case NSs might feel at a disadvantage in ELF contexts, which is indeed pointed out sometimes at least anecdotally.

There are other research questions the concept of approximation can generate. For example, the forms that approximations of multi-word units can take in L2 use can give insight into the levels of abstraction at which chunking or syntagmatic prospection can function or shed light on the cognitive constraints of multilingual processing, being their primary outcome.

Further, it was hypothesised in this study that the mechanism of the idiom principle is based on syntagmatic association and implicit memory. Some properties of this syntagmatic association, like the impact of verbatimness, contiguity and the direction of association on its strength, were also tentatively suggested. But there are many more questions to ask. For example, the role of meaning in formation and subsequent entrenchment of syntagmatic association remained unclear. Though theoretical analysis of the question predicts that it is the common communicative function which unites the components of a unit of meaning and gives rise to the emergence of syntagmatic association, empirical observations do not entirely corroborate this assumption. At least they suggest that syntagmatic association is not impossible outside a unit of meaning, which implies that mere proximity might play a bigger role in the formation of syntagmatic association. It is possible that both proximity and meaning are important. If so, it would be interesting to find out how they interact, a goal which seems to be in line with the objectives of research into language emergence and language as a complex adaptive system (see Beckner et al. 2009; Ellis and Larsen-Freeman 2006).

Indeed, this brings us to the idea that in natural language use units of different underlying structures may be colliding and interacting with each other, as a result of which their configurations might be changing or new units appearing. For example, Linear Unit Grammar (Sinclair and Mauranen 2006) proposes a linear model of chunking which attempts to describe how we naturally chunk incoming language without resorting to complex hierarchies and starting with an intuitive, pre-theoretical assignment of chunk boundaries which makes linear chunks plausible candidates for the role of units of processing. However, the boundaries of a linear chunk do not necessarily correspond with the boundaries of a unit of meaning. Thus, it would be interesting to find out what the relationship between linear units of processing and units of meaning is: How do they interact? Does the interaction eventually result in equilibrium? Also, what are other levels at which syntagmatic prospection works, all the way up to the context of situation, and, how do they influence the production, comprehension and formation of units of meaning and units of processing? In addition to this, we also have the interaction of primings from different domains of language use.

All these levels of interaction and the ensuing potential for variation and change suggest that in the long run we might have to abandon the notion of lexical storage as it positions lexis as too static and does not reflect its capacity for dynamic development and variability.[106] The lexicon might indeed be empty after all (Sinclair 1996b).

[106] For example, Ellis (2008) suggests that development in cognitive neuroscience encourages "a shift of emphasis from knowledge as static representation stored in particular locations to knowledge as processing involving the dynamic mutual influence of interrelated types of information as they activate and inhibit each other over time" (6). See also Ellis (2012a: 21-22).

References

Ädel, Annelie. 2006. *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamins.

Albrechtsen, Dorte, Kirsten Haastrup & Birgit Henriksen. 2008. *Vocabulary and writing in a first and second language: Processes and development*. Basingstoke; New York: Palgrave Macmillan.

Altenberg, Bengt. 1998. On the phraseology of spoken English: The evidence of recurrent word combinations. In Anthony P. Cowie (ed.), *Phraseology: Theory, analysis and applications,* 101-122. Oxford: Clarendon.

Anthony, Laurence. 2007. *AntConc (Version 3.2.4w)* [Software]. Available at: http://www.antlab.sci.waseda.ac.jp/software.html (last accessed 8.4.2013).

Beckner, Clay, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman & Tom Schoenemann. 2009. Language is a complex adaptive system. Position paper, *Language Learning* 59, Supplement 1. 1-27.

Biber, Douglas. 1988. *Variation across speech and writing.* Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *The Longman grammar of spoken and written English.* London: Longman.

Bock, Kathryn J. & William F. Brewer. 1974. Reconstructive recall in sentences with alternative surface structures. *Journal of Experimental Psychology* 103(5). 837-843.

Bolinger, Dwight. 1976. Meaning and memory. *Forum Linguisticum* 1(1). 1-14.

BNC: The British National Corpus. http://bncweb.lancs.ac.uk/

Britton, Bruce K. 1994. Understanding expository text: Building mental structure to induce insights. In Morton A. Gernsbacher (ed.), *Handbook of psycholinguistics*, 641-674. New York: Academic Press.

Bybee, Joan. 2002. Sequentiality as the basis of constituent structure. In Talmy Givón & Bertram Malle (eds.), *The evolution of language from pre-language*, 109-32. Amsterdam: John Benjamins.

Bybee, Joan. 2010. *Language, usage and cognition.* Cambridge: Cambridge University Press.

Carey, Ray. 2013. On the other side: Formulaic organizing chunks in spoken and written academic ELF. *Journal of English as a Lingua Franca* 2(2). 207–228

Carter, Ronald & Michael McCarthy. 2006. *Cambridge grammar of English: A comprehensive guide.* Cambridge: Cambridge University Press.

Cheng, Winnie, Chris Greaves & Martin Warren. 2006. “From n-gram to skipgram to concgram”. *International Journal of Corpus Linguistics* 11(4). 411–433.

Cheng, Winnie, Chris Greaves, John McH. Sinclair & Martin Warren. 2009. Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics* 30(2). 236–252.

Clark, Herbert H. 1970. Word associations and linguistic theory. In John Lyons (ed.), *New horizons in linguistics*, 271–286. Harmondsworth: Penguin.

Clear, Jeremy. 1993. From Firth principles: Computational tools for the study of collocation. In Mona Baker, Gill Francis & Eelena Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair,* 271-292. Amsterdam: John Benjamins.

COCA: The Corpus of Contemporary American English. http://corpus.byu.edu/coca/

Cowie, 1981. The treatment of collocations and idioms in learners’ dictionaries. *Applied Linguistics* 2(3). 223-235.

Cowie 1998. Introduction. In Anthony P. Cowie (ed.), *Phraseology: Theory, analysis, and applications*, 1-20. Oxford: Clarendon.

Coxhead, Averil. 2000. A new Academic Word List. *TESOL Quarterly* 34. 213-238.

Dąbrowska, Eva. 2004. *Language, mind and brain*. Edinburgh, Scotland: Edinburgh University Press.

De Cock, Sylvie. 2000. Repetitive phrasal chunkiness and advanced EFL speech and writing. In Christian Mair & Marianne Hundt (eds.), *Corpus linguistics and linguistic theory*, 51-68. Amsterdam: Rodopi.

Deese, James. 1965. *The structure of associations in language and thought*. Baltimore: The Johns Hopkins Press.

EAT: Edinburgh Associative Thesaurus. http://www.eat.rl.ac.uk/ (last accessed 4.12.2013).

ELFA corpus. The corpus of English as a lingua franca in academic settings. Director: Anna Mauranen. http://www.helsinki.fi/elfa/elfacorpus

Ellis, Nick C. 1996. Sequencing in SLA: Phonological memory, chunking and points of order. *Studies in Second Language Acquisition* 18. 91-126.

Ellis, Nick C. 2002. Frequency effects in language acquisition: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24. 143-188.

Ellis, Nick C. 2006. Cognitive perspectives on SLA: The Associative Cognitive CREED. *AILA Review* 19. 100–121.

Ellis, Nick C. 2009. Optimizing the input: Frequency and sampling in usage-based and form-focused learning. In Michael H. Long & Catherine J. Doughty (eds.), *The handbook of language teaching*, 139- 158. Malden, MA: Wiley-Blackwell.

Ellis, Nick C. 2011. The emergence of language as a complex adaptive system. In James Simpson (ed.), *The Routledge handbook of applied linguistics*. 666–679. London: Routledge.

Ellis, Nick C. 2012a. What can we count in language, and what counts in language acquisition, cognition, and use? In S. Th. Gries & D. S. Divjak (Eds.) *Frequency effects in language learning and processing (vol. 1),* 7-34. Berlin: Mouton de Gruyter.

Ellis, Nick C. 2012b. Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics* 32. 17-44.

Ellis, Nick C. & Fernando Ferreira–Junior. 2009. Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal* 93(3). 370–385.

Ellis, Nick C. & Diane Larsen-Freeman. 2006. Language emergence: Implications for Applied Linguistics. [Introduction to the special issue]. *Applied Linguistics* 27(4). 558–589.

Ellis, Nick C. & Eric Frey. 2009. The psycholinguistic reality of collocation and semantic prosody (2): Affective priming. In Roberta Corrigan, Edith Moravcsik, Hamid Ouali & Kathleen Wheatley (eds.), *Formulaic language (vol. 2): Acquisition, loss, psychological reality, and functional explanations* [Typological Studies in Language, 83], 473-497. Amsterdam: John Benjamins.

Ellis, Nick C., Eric Frey, & Isaac Jalkanen. 2009. The psycholinguistic reality of collocation and semantic prosody (1): Lexical access. In Ute Römer & Rainer Schulze (eds.), *Exploring the lexis-grammar interface* [Studies in Corpus Linguistics, 35], 89-114. Amsterdam: John Benjamins.

Ellis, Nick C., Matthew Brook O'Donnell & Ute Römer. 2013. Usage-based language: investigating the latent structures that underpin acquisition. *Language Learning* 63. 25–51.

Ellis, Nick C., Matthew Brook O'Donnell & Ute Römer. 2014. The processing of verb-argument constructions is sensitive to form, function, frequency, contingency and prototypicality. *Cognitive Linguistics* 25(1). 55-98.

Erman, Britt & Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text* 20(1). 29-62.

Erman, Britt. 2009. Formulaic language from a learner perspective: What the learner needs to know. In Roberta Corrigan et al. (eds.), *Formulaic language (vol. 2): Acquisition, loss, psychological reality, and functional explanations,* 323-346. Amsterdam : John Benjamins.

Evert, Stefan. 2004. The statistics of word co-occurrences: Word pairs and collocations. PhD dissertation, University of Stuttgart.

Evert, Stefan. 2008. Corpora and collocations. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An international handbook*, 1212-1248. Berlin: Mouton de Gruyter.

Firth, John R. 1957. Modes of meaning. In John R. Firth, *Papers in Linguistics 1934 – 1951*, 190-215. London: Oxford University Press.

Firth, John R. 1968 [1957]. A synopsis of linguistic theory, 1930 -1955. In Frank R. Palmer (ed.), *Selected papers of J.R. Firth 1952-1959,* 168-205. London: Longman

Fitzpatrick, Tess. 2006. Habits and rabbits: Word associations and the L2 lexicon. *EUROSLA Yearbook* 6. 121-145.

Fitzpatrick, Tess. 2007. Word association patterns: Unpacking the assumptions. *International Journal of Applied Linguistics* 17(3). 319-331.

Fitzpatrick, Tess. 2009. Word association profiles in a first and second language: Puzzles and problems. In Tess Fitzpatrick & Andy Barfield (eds.), *Lexical processing in second language learners: Papers and perspectives in honour of Paul Meara,* 38-52. Bristol: Multilingual Matters.

Fitzpatrick, Tess. 2013. Word associations. In Carol A. Chapelle (ed.), *Encyclopedia of Applied Linguistics*, 6193–6199. Wiley-Blackwell. (accessed online 12.11.2013).

Fletcher, William H. Phrases in English. http://phrasesinenglish.org/ (last accessed 23.12. 2013).

Fletcher, William. 2002-2012. *kfNgram*. http://www.kwicfinder.com/kfNgram/kfNgramHelp.html (last accessed 8.10.2013).

Fox, Gwyneth.1987. The case for examples. In John McH. Sinclair (ed.), *Looking up: An account of the COBUILD project in lexical computing,* 137-149. London: Collins.

Francis, Gill, Susan Hunston & Elizabeth Manning. 1998. *Collins COBUILD Grammar Patterns: Nouns and adjectives*. London: HarperCollins.

Gahl, Susanne & Alan C. L. Yu. 2006. Introduction to the special issue on exemplar-based models in linguistics. *The Linguistic Review 23.* 213–216

Galton, Francis. 1879. Psychometric experiments. *Brain* 2. 149–162.

Geeraerts, Dirk. 2010. *Theories of lexical semantics.* Oxford: Oxford University Press.

Gilquin, Gaëtanelle, Granger, Sylviane, & Paquot, Magali. 2007. Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes* 6. 319-335.

Gomez, Rebecca L. & LouAnn Gerken 1999. Artificial grammar learning by 1-year olds leads to specific and abstract knowledge. *Cognition* 70. 109 – 35.

Gomez, Rebecca L. & LouAnn Gerken. 2000. Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences* 4. 178 - 86.

Gomez, Rebecca L. 2001. Finding structure in language: Sometimes more variability is better. Presentation given at the annual meeting of the AAAS in the symposium: Tools Infants Might Use to Learn Language. San Francisco, February.

Graddol, David. 2006. *English next. Why global English may mean the end of "English as a foreign language"*. London: British Council.

Gramley, Stephan & Kurt-Michael Pätzold. 1992. *A survey of modern English*. London: Routledge.

Granger, Sylviane. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In Karin Aijmer, Bengt Altenberg & Mats Johansson (eds.), *Languages in contrast: Papers from a symposium on text-based cross-linguistic studies, Lund 4-5 March 1994*, 37-51. Lund: Lund University Press.

Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In Anthony P. Cowie (ed.), *Phraseology: Theory, analysis, and applications,* 145-160. Oxford: Clarendon.

Granger, Sylviane 2002. A Bird's-eye view of leaner corpus research. In Sylviane Granger, Joseph Hung, & Stephanie Petch-Tyson (eds.), *Computer learner corpora, second language acquisition and foreign language teaching*, 3-33. Amsterdam: John Benjamins.

Granger, Sylviane. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In Karin Aijmer (ed.), *Corpora and language teaching*, 13-31. Amsterdam: John Benjamins.

Granger, Sylviane & Fanny Meunier. 2008. Phraseology in language learning and teaching: Where to from here? In Fanny Meunier & Sylviane Granger (eds.), *Phraseology in foreign language learning and teaching*, 247-252. Amsterdam: John Benjamins.

Granger, Sylviane & Magali Paquot. 2008. Disentangling the phraseological web. In Sylviane Granger & Fanny Meunier (eds.), *Phraseology. An interdisciplinary perspective*, 27- 49. Amsterdam: John Benjamins.

Granger, Sylviane & Magali Paquot. 2009. Lexical verbs in academic discourse: A corpus-driven study of learner use. In Maggie Charles, Diane Pecorari & Susan Hunston (eds.), *Academic writing: At the interface of corpus and discourse*, 193-214. New York: Continuum.

Greaves, Chris. 2009 *ConcGram 1.0: A phraseological search engine*. [Software]. Amsterdam: John Benjamins.

Gurevich, Olga, Matthew A. Johnson & Adele E. Goldberg. 2010. Incidental verbatim memory for language. *Language and Cognition* 2(1). 45-78.

Hiltunen, Turo. 2010 *Grammar and disciplinary culture: A corpus-based study*. Helsinki: University of Helsinki doctoral dissertation. Available at: http://urn.fi/URN:ISBN:978-952-10-6464-7

Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. London: Routledge.

Hoey, Michael. 2009. Corpus-driven approaches to grammar: The search for common ground. In Ute Römer & Rainer Schulze (eds.), *Exploring the lexis-grammar interface,* 33-47. Amsterdam; Philadelphia: John Benjamins.

Hoey, Michael & Matthew Brook O'Donnell. 2008. Lexicography, grammar, and textual position. *International Journal of Lexicography* 21(3). 293-309.

Howarth, Peter. 1998. The Phraseology of learners' academic writing. In Anthony P. Cowie (ed.), *Phraseology: Theory, analysis, and applications,* 161-186. Oxford: Clarendon.

Hu, Marcella & Paul Nation. 2000. Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language* 13(1). 403-430.

Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language.* Cambridge: Cambridge University Press.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, Susan. 2007. Semantic prosody revisited. *International Journal of Corpus Linguistics* 12(2). 249-268.

Hunston, Susan. 2008. Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics* 13(3). 271-295.

Hunston, Susan. 2010. Starting with the small words. In Ute Römer & Rainer Schulze (eds.), *Patterns, meaningful units and specialized discourses*, 7-30. Amsterdam; Philadelphia: John Benjamins.

Hunston, Susan. 2011. *Corpus approaches to evaluation.* New York: Routledge.

Hunston, Susan & Gill Francis. 2000. *Pattern grammar.* Amsterdam: John Benjamins.

Hyland, Ken. 2005. *Metadiscourse: Exploring interaction in writing.* London: Continuum.

Hyland, Ken. 2008. Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics* 18(1). 41-62.

Hynninen, Niina. 2013. *Language regulation in English as a lingua franca: Exploring language-regulatory practices in academic spoken discourse.* Helsinki: University of Helsinki doctoral dissertation. Available at: http://urn.fi/URN:ISBN:978-952-10-8639-7

ICLE corpus. The International Corpus of Learner English. http://www.uclouvain.be/cecl-icle.html

Irujo, Suzanne. 1986. A piece of cake: Learning and teaching idioms. *ELT Journal* 40. 236-242.

Irujo, Suzanne. 1993. Steering clear: Avoidance in the production of idioms. *International Review of Applied Linguistics in Language Teaching* 31. 205- 219.

Jarvis, Scott. 2000. Methodological rigor in the study of transfer: Identifying L1 influence in the interlanguage lexicon. *Language Learning* 50 (2). 245-309.

Jung, Carl G. 1910. The association method. *American Journal of Psychology* 21. 219–269.

Kaszubski Przemek. 2000. *Selected aspects of lexicon, phraseology and style in the writing of Polish advanced learners of English: A contrastive, corpus-based.* PhD Thesis. Poznań: Adam Mickiewicz University.

Kellerman, Eric. 1978. Giving learners a break: Native language intuitions as a source of predictions about transferability. *Working Papers in Bilingualism* 15. 309-315.

Kent, Grace Helen & A. J. Rosanoff. 1910. A study of association in insanity. Part I. *American Journal of Insanity* 67(1). 37–96.

Kiss, G.R. 1968. Words, associations and networks. *Journal of Verbal Learning and Verbal Behaviour* 7. 707-713.

Kiss, G.R., C. Armstrong, R. Milroy & J. Piper. 1973. An associative thesaurus of English and its computer analysis. In A.J. Aitken, R.W. Bailey & N. Hamilton-Smith (eds.), *The computer and literary studies*, 153-165. Edinburgh: University Press.

Kjellmer Goran. 1991. A mint of phrases. In Karin Aijmer & Bengt Altenberg (eds.), *English corpus linguistics: Studies in honour of Jan Svartvik,* 111-127. London: Longman.

Laufer, Batia. 2000. Avoidance of idioms in a second language: The effect of L1 - L2 degree of similarity. *Studia Linguistica* 54. 186-196.

Laufer, Batia & Stig Eliasson. 1993. What causes avoidance in L2 1earning: Ll - L2 difference, L1 - L2 similarity, or L2 complexity? *Studies in Second Language Acquisition* 15. 35-48.

Laufer, Batia & Paul Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics* 16(3). 307-322.

Lave, Jean & Etienne Wenger. 1991. *Situated learning: Legitimate peripheral participation.* Cambridge: Cambridge University Press.

Leech, Geoffrey. 1998. Preface. In Sylviane Granger (ed.), *Learner English on computer*, xiv-xx. London and New York: Longman

Lehtonen, Minna & Matti Laine. 2003. How word frequency affects morphological processing in monolinguals and bilinguals. *Bilingualism: Language and Cognition* 6(3). 213–225

Levelt, Willem J. M. 2013. *A history of psycholinguistics: The pre-Chomskyan era*. Oxford: Oxford University Press.

Louw, Bill. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair,* 157-176. Amsterdam: John Benjamins.

Louw, Bill. 2000. Contextual prosodic theory: Bringing semantic prosodies to life. In Chris Heffer & Helen Sauntson (eds.), *Words in context: A tribute to John Sinclair on his retirement*, 48-94. English Language Research Discourse Analysis Monograph no. 18, CD-ROM. Birmingham, AL: University of Birmingham. http://www.revue-texto.net/docannexe/file/124/louw_prosodie.pdf (last accessed 29.1.2014).

Malkiel, Yakov. 1959. Studies in irreversible binomials. *Lingua* 8. 113-160.

Mauranen, Anna. 1993. *Cultural differences in academic rhetoric*. Frankfurt am Main: Peter Lang.

Mauranen, Anna. 2004. Where Next? A summary of the round table discussion. In Gabriella Del Lungo Camiciotti & Elena Tognini-Bonelli (eds.), *Academic discourse: new insights into evaluation*, 203-216. Bern: Peter Lang.

Mauranen, Anna. 2005. English as a Lingua Franca—an unknown language? In Giuseppina Cortese & Anna Duszak (eds.*), Identity, community, discourse: English in intercultural settings*, 269–293. Frankfurt: Peter Lang.

Mauranen, Anna. 2009. Chunking in ELF: Expressions for managing interaction. *Journal of Intercultural Pragmatics* 6(2). 217-233.

Mauranen, Anna. 2010. Features of English as a lingua franca in academia. *Helsinki English Studies* 6. 6–28.

Mauranen, Anna. 2011. Learners and users – Who do we want corpus data from? In Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin & Magali Paquot (eds.), *A taste for corpora: In honour of Sylviane Granger*, 155-171. Amsterdam: John Benjamins.

Mauranen, Anna. 2012. *Exploring ELF: Academic English shaped by non-native speakers*. Cambridge: Cambridge University Press.

McEnery, Tony, Richard Xiao & Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book.* London and New York: Routledge.

Meara, Paul. 1983. Word associations in a second language. *Nottingham Linguistics Circular* 11. 28–38.

Meara, Paul. 1997. Towards a new approach to modelling vocabulary acquisition. In Norbert Schmitt & Michael McCarthy (eds.), *Vocabulary: Description, acquisition and pedagogy,* 109-121. Cambridge: Cambridge University Press.

Meara, Paul. 2009. *Connected words: Word associations and second language vocabulary acquisition*. Amsterdam; Philadelphia: John Benjamins.

Meunier, Fanny & Sylviane Granger (eds.). 2008. *Phraseology in foreign language learning and Teaching*. Amsterdam: John Benjamins.

MICASE corpus. The Michigan Corpus of Academic Spoken English. http://quod.lib.umich.edu/m/micase/

Michelbacher, Lukas, Stefan Evert & Hinrich Schütze. 2011. Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory* 7(2). 245‑276.

Mittwoch, Anita, Rodney Huddleston & Peter Collins. 2002. The clause: Adjuncts. In Rodney Huddleston & Geoffrey K. Pullum (eds.), *The Cambridge grammar of the English language,* 663-784. Cambridge, UK: Cambridge University Press.

Mollin, Sandra. 2009. Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory* 5(2). 175–200.

Mollin, Sandra. 2011. Order and law? Degrees of reversibility in English binomials. Paper presented at ICAME 32: Trends and Traditions in English Corpus Linguistics, in Honour of Stig Johansson, Oslo, June 1-5.

Morley, John & Alan Partington. 2009. A few frequently asked questions about semantic or evaluative prosody. *International Journal of Corpus Linguistics* 14(20). 139-158.

Morris, Charles W. 1938. Foundations of the theory of signs. In Otto Neurath, Rudolph Carnap & Charles W. Morris (eds.), *International encyclopedia of unified science*, 77-138. Chicago: Chicago University Press.

Moss, Helen & Lianne Older. 1996. *Birkbeck word association norms*. Hove: Psychology Press.

Nation, Paul. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, Paul. 2006. How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review* 63(1). 59-82

Nesselhauf, Nadja. 2004. What are collocations? In David Allerton, Nadja Nesselhauf, Paul Skandera (eds.), *Phraseological units: Basic concepts and their application*, 1-21. Basel: Schwabe.

Nesselhauf, Nadja. 2005. *Collocations in a learner corpus*. Amsterdam: Benjamins.

Oxford Dictionary of English, 3edn. 2010. Oxford: Oxford University Press.

Palermo David S. & James J. Jenkins 1964. *Word association norms*. Minneapolis: University of Minnesota Press.

Paquot, Magali. 2007. *EAP vocabulary in EFL learner writing: From extraction to analysis: A phraseology-oriented approach.* Unpublished PhD thesis. Universite catholique de Louvain, Centre for English Corpus Linguistics.

Paquot, Magali. 2008. Exemplification in learner writing: A cross-linguistic perspective. In Meunier, Fanny & Granger, Sylviane (eds.), *Phraseology in foreign language learning and teaching,* 101- 119. Amsterdam: John Benjamins.

Paquot, Magali. 2010. *Academic vocabulary in learner writing: From extraction to analysis*. London & New-York: Continuum.

Paribakht, Sima T. & Marjorie Bingham Wesche. 1993. Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal* 11 (1). 9-29

Partington, Alan. 1998. *Patterns and meanings: Using corpora for English language research and teaching*. Amsterdam: Benjamins.

Partington, Alan. 2004. "Utterly content in each other's company": Semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9(1). 131-156.

Partington, Alan. 2006. *The linguistics of laughter: A corpus-assisted study of laughter talk.* London; New York: Routledge.

Pawley, Andrew & Frances H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Jack C. Richards & Richard W. Schmidt (eds.), *Language and communication*, 191–227. London: Longman.

PDFX v1.8 [Software]. Available at: http://pdfx.cs.man.ac.uk/ (last accessed 28.10.2013).

Peters, Ann M. 1983. *The unit of language acquisition.* Cambridge: Cambridge University Press.

Philip, Gill. 2011. *Colouring meaning: Collocation and connotation in figurative language.* Amsterdam; Philadelphia: John Benjamins.

Pitzl, Marie-Luise. 2009. "We should not wake up any dogs": Idiom and metaphor in ELF. In Anna Mauranen & Elina Ranta (eds.), *English as a lingua franca: Studies and findings,* 298-322. Newcastle upon Tyne: Cambridge Scholars Press.

Pitzl, Marie-Luise. 2012. Creativity meets convention: idiom variation and re-metaphorization in ELF. *Journal of English as a Lingua Franca* 1(1). 27-55.

Pollio, H.R. 1966. *The structural basis of word association behaviour*. The Hague: Mouton.

Postman, Leo & Geoffrey Keppel (eds.). 1970. *Norms of word association*. New York : Academic Press.

Poston Dudley L. & Leon F. Bouvier. 2010. *Population and society: An introduction to demography.* Cambridge: Cambridge University Press.

Pullum, Geoffrey K. & Rodney Huddleston 2002. Prepositions and prepositional phrases. In Rodney Huddleston and Geoffrey K. Pullum, *The Cambridge grammar of the English language*, 597-661. Cambridge: Cambridge University Press

Renouf, Antoinette. 1987. Moving on. In John McH. Sinclair (ed.), *Looking up: An account of the COBUILD project in lexical computing,* 167-178. London: Collins.

Römer, Ute. 2011. Observations on the phraseology of academic writing: Local patterns – local meanings? In Thomas Herbst, Susen Faulhaber & Peter Uhrig (eds.), *The phraseological view of language: A tribute to John Sinclair*, 211-227. Berlin: Mouton de Gruyter.

Saffran, Jenny R., Richard N. Aslin & Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274. 1926-1928.

Schmitt, Norbert. 1998a. Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning* 48(2). 281–317

Schmitt, Norbert. 1998b. Quantifying word association responses: What is native-like? *System 26.* 389–401.

Schmitt, Norbert (ed.). 2004. *Formulaic sequences: Acquisition, processing and use*. Amsterdam: John Benjamins

Schmitt, Norbert. 2010. *Researching vocabulary: A vocabulary research manual.* Basingstoke: Palgrave Macmillan.

Schmitt, Norbert & Ronald Carter. 2004. Formulaic sequences in action: An introduction. In Norbert Schmitt (ed.), *Formulaic sequences: Acquisition, processing and use,* 1-22. Amsterdam: John Benjamins

Schmitt, Norbert, Sarah Grandage & Svenja Adolphs. 2004. Are corpus-derived recurrent clusters psycholinguistically valid? In Norbert Schmitt (ed.), *Formulaic sequences: Acquisition, processing, and use,* 127–151. Amsterdam & Philadelphia: Benjamins.

Scott Mike & Christopher Tribble. 2006. *Textual patterns: Key words and corpus analysis in language education.* Amsterdam: John Benjamins.

Segalowitz, Norman & Jan Hulstijn. 2009. Automaticity in bilingualism and second language learning. In Judith F. Kroll & Annette M.B. de Groot (eds.), *Handbook of bilingualism*, 371- 388. New York, NY: Oxford University Press.

Seidlhofer, Barbara. 2009. Accommodation and the idiom principle in English as a Lingua Franca. *Intercultural Pragmatics* 6(2). 195–215.

Simpson-Vlach, Rita & Nick C. Ellis. 2010. An Academic Formulas List (AFL). *Applied Linguistics* 31. 487-512.

Sinclair, John McH. 1981. Planes of discourse. In S.N.A. Rizvi (ed.), *The two-fold voice: Essays in honour of Ramesh Mohan*, 70-91. India: Pitambar Publishing.

Sinclair, John McH. 1987. Collocation: a progress report. In Ross Steele & Terry Treadgold (eds.), *Language topics: Essays in honour of Michael Halliday,* 319-331. Amsterdam: John Benjamins.

Sinclair, John McH. (ed.). 1990. *Collins COBUILD English grammar.* London: HarperCollins.

Sinclair, John McH. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, John McH. 1996a. The search for units of meaning. *Textus* 9 (1). 75-106.

Sinclair, John McH. 1996b. The empty lexicon. *International Journal of Corpus Linguistics* 1 (1). 99-119.

Sinclair, John McH. 1998. The lexical item. In Edda Weigand (ed.), *Contrastive lexical semantics*, 1-24. Amsterdam: John Benjamins.

Sinclair, John McH. 2001. A tool for text explication. In Karin Aijmer (ed.), A *wealth of English: Studies in honour of Göran Kjellme*r. Göteborg: Acta Universitatis Gothoburgensis.

Sinclair, John McH. 2004. *Trust the text.* London: Routledge.

Sinclair, John McH. 2007. Collocation reviewed. [Manuscript]. Tuscan Word Centre, Italy.

Sinclair, John McH. 2008. The phrase, the whole phrase and nothing but the phrase. In Sylviane Granger & Fanny Meunier (eds.), *Phraseology. An interdisciplinary perspective*, 407- 410. Amsterdam: John Benjamins.

Sinclair, John McH, Susan Jones, Robert Daley & Ramesh Krishnamurthy. 2004. *English collocation studies: The OSTI report.* [Including a new interview with John Sinclair conducted by Wolfgang Teubert]. London: Continuum.

Sinclair, John McH. & Anna Mauranen. 2006. *Linear unit grammar.* Amsterdam: John Benjamins.

Sosa, Anna Vogel & James MacFarlane. 2002. Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word of. *Brain and Language* 83(2). 227-236.

Stewart, Dominic. 2010. *Semantic prosody: A critical evaluation.* New York: Routledge.

Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language* 2(1). 23-55.

Stubbs, Michael. 2001. *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

Stubbs, Michael. 2007. Quantitative data on multi-word sequences in English: The case of the word *world*. In Michael Hoey, Michaela Mahlberg, Michael Stubbs & Wolfgang Teubert (eds.), *Text, discourse and corpora: Theory and analysis,* 163-189. London and New York: Continuum.

Stubbs, Michael. 2009. Memorial article: John Sinclair (1933-2007). The search for units of meaning: Sinclair on empirical semantics. *Applied Linguistics* 30(1). 115-137.

Stubbs, Michael. 2011. Sequence and order: The neo-Firthian tradition of corpus semantics. Paper presented at ICAME 32: Trends and Traditions in English Corpus Linguistics, in Honour of Stig Johansson, Oslo, June 1-5.

Swales, John. 1990. *Genre analysis: English in academic and research settings*. New York: Cambridge University Press.

Teubert, Wolfgang. 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics* 10(1). 1–13.

Teubert, Wolfgang. 2010. *Meaning, discourse and society.* Cambridge: Cambridge University Press.

Thompson, Geoff & Susan Hunston. 2000. Evaluation: An introduction. In Susan Hunston & Geoff Thompson (eds.), *Evaluation in text: Authorial stance and the construction of discourse*, 1-27. Oxford: Oxford University Press.

Thumb, Albert & Karl Marbe. 1901. *Experimentelle Untersuchungen uÈber die psychologischenGrundlagen der sprachlichen Analogiebildungen*. Leipzig: Engelmann.

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work.* Amsterdam: John Benjamins.

Tomasello, Michael. 2003. *Constructing a language.* Cambridge, MA: Harvard University Press.

Trautscholdt, Martin. 1883. Experimentelle Untersuchungen über die Association der Vorstellungen. *Philosophische Studien 1*. 213-250.

Vihman, Marilyn M. 1982. Formulas in first and second language acquisition. In Loraine K. Obler & Lise Menn (eds.), *Exceptional language and Linguistics*, 261-284. New York: Academic Press.

Warren, Martin. 2010. Identifying aboutgrams in engineering texts. In Marina Bondi & Mike Scott (eds), *Keyness in texts*, 113-126. Amsterdam: John Benjamins.

Werlich, Egon. 1976. *A text grammar of English*. Heidelberg: Ouelle and Meyer.

West, Michael. 1953. *A general service list of English words.* London: Longman, Green & Co.

Whitsitt, Sam. 2005. A critique of the concept of semantic prosody. *International Journal of Corpus Linguistics* 10(3). 283-305.

Wolter, Brent. 2002. Assessing proficiency through word associations: Is there still hope? *System* 30(3). 315–329.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, Alison. 2008. *Formulaic language: Pushing the boundaries.* Oxford: Oxford University Press.

Wray, Alison. 2009. Identifying formulaic language: Persistent challenges and new opportunities. In Roberta Corrigan et al. (eds.), *Formulaic language, vol.1: Distribution and historical change,* 27-51. Amsterdam: John Benjamins.

Wray, Alison & Kazuhiko Namba. 2003. Formulaic language in a Japanese-English bilingual child: A practical approach to data analysis. *Japan Journal for Multilingualism and Multiculturalism* 9(1). 24-51.

Yorio, Carlos A. 1989. Idiomaticity as an indicator of second language proficiency. In Kenneth Hyltenstam & Loraine K. Obler (eds.) *Bilingualism across the lifespan*, 55-72. Cambridge: Cambridge University Press.

Zipf, George K. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: MIT Press.

Appendix A: A sample word association task

Please write down the first word (s) you think of when you read each of the words listed, as fast as you can.

optimality	
proto	
strictly	
paper	
seem	
instance	
closely	
general	
similarity	
significant	
each	
strong	
whether	
automatic	
example	
rewrite	
early	
orthographic	
mean	
possible	
caused	
sentence	
use	
concentrate	
maintain	
amount	
computational	
areas	
set	

every	
roughly	
benefits	
calculated	
evidence	
stem	
coverage	
such	
different	
perceived	
even	
proves	
scarcity	
have	
approach	
initial	
establish	
concise	
look	
machine	
since	
makes	
surface	
account	
rely	
like	
analysis	
related	
along	
solution	
take	
most	

order	
so	
character	
dealing	
comparative	
followed	
word	
basic	
lot	
false	
descend	
theoretically	
sake	
training	
TWiki	
phonemic	
case	
distant	
geminated	
affinity	
mark	
other	
family	
suggests	
version	
few	
results	
intend	
refine	
mapped	
already	
purpose	

rules	
specifically	
manually	
original	
traditionally	
lack	
provided	
efficient	
unnoticed	
for	
also	
refer	
whereas	
recognition	
number	
systematic	
attached	
variation	
knowledge	
crucial	
bears	
insufficient	
hard	
illustrate	
there	
bilingual	
restricting	
linguistics	

Appendix B: C1 concgrams compared to C2 (Maisa and Kaisa)

Table B.1 Maisa, matching concgrams

Co-occurring word	Co-occurring word	C1	C2	Pattern represented by the concgram
AS	WELL	51	112	*(as) well as*
HIV/AIDS	AIDS	48	70	*HIV/AIDS*
BEEN	HAS	42	61	*has been*
BEEN	HAVE	36	76	*have been*
EXAMPLE	FOR	32	69	*for example*
AL	ET	31	14	*et al*
EXPECTANCY	LIFE	27	19	*life expectancy*
FERTILITY	PREMARITAL	26	107	*premarital fertility*
BASED	ON	26	80	*based on*
AIDS	DEATHS	25	22	*AIDS deaths, AIDS-related deaths, deaths due to AIDS*
AREAS	URBAN	24	51	*urban areas, urban and rural areas*
AREAS	RURAL	24	50	*rural areas*
AIDS	IMPACT	24	14	*AIDS impact model, the impact of (HIV/) AIDS*
AVAILABLE	DATA	23	57	*available (HIV prevalence, surveillance) data, data (is/are/were /becomes,* etc.) *available*
EVEN	THOUGH	19	4	*even though*
DETERMINANTS	PROXIMATE	19	2	*proximate determinants (of fertility)*
PREGNANT	WOMEN	18	143	*pregnant women*
NO	THERE	18	22	*there* BE *no*
AFRICA	SUB	18	9	*sub-Saharan Africa*

GROWTH	RATE	18	8	*(population) growth rate, rate of growth of the epidemic*
SUB-SAHARAN	AFRICA	17	7	*sub-Saharan Africa*
HIV/AIDS	IMPACT	15	7	*the impact of HIV/AIDS*
ANTIRETROVIRAL	TREATMENT	14	3	*antiretroviral treatment*
IN	SUB	13	95	C1: *in S/sub-Saharan Africa,* C2, e.g.: *in sub-Saharan Africa, in sub-populations*
ASSUMPTIONS	ABOUT	13	27	*assumptions about* (e.g. *HIV, fertility)*
DOES	NOT	13	25	*does not*
IT	POSSIBLE	13	19	*it is (not) possible to* + inf./*that*
RATE	TOTAL	13	5	*total fertility rate* [abbreviated *TFR*]
CAN	SEEN	13	1	*can* BE *seen (in)*
FIRST	MARRIAGE	12	59	*first marriage* (e.g. *age at first marriage)*
MORTALITY	MIGRATION	12	11	*(levels of) fertility, mortality and migration*
NOT	YET	12	5	auxiliary verb/BE+ *not yet* + past participle/adj.
MEN	WOMEN	11	63	*men and women, women and men*
HIGHER	THAN	11	33	*higher than* [used to talk about *prevalence, incidence, number of* something]
AT	BIRTH	11	26	C1*: at birth,* e.g. *life expectancy, age, sex ratio at birth* C2*: at (first) birth,* e.g. *mother's age, life expectancy, size, sex ratio at birth*
BIRTHS	NUMBER	11	6	*number of (live/total) births*
AIDS	WITHOUT	11	3	*without AIDS* C2: *“Without-AIDS Scenario”*
AGE	STRUCTURE	10	20	*age structure* (e.g. *of deaths, infection, the population)*
LOWER	THAN	10	7	*(X* BE*) lower than (Y)*
ANALYSIS	UNCERTAINTY	10	7	*uncertainty analysis (for HIV prevalence)*

DO	NOT	9	36	*do not*
ART	COVERAGE	9	31	C1&C2*: ART coverage,* C2 also: *coverage of ART*
CONTRACEPTION	USE	9	10	*use of (modern) contraception, contraception use,* USE *contraception*
MEANS	THIS	9	5	*this means (that/*NP/clause*)*
AMONG	WOMEN	8	124	*among (young/er, married, pregnant* etc.) *women*
SENTINEL	SITES	8	55	*sentinel (survey/surveillance) sites*
AGE	DISTRIBUTION	8	33	*age (and sex) distribution*
SOFTWARE	SPECTRUM	8	11	*Spectrum software, software called Spectrum*
BY	USING	8	10	*by using* X
FACT	THAT	8	8	*the fact that*
FUTURE	TRENDS	8	6	C1*: future (population) trends,* C2*: future (fertility/mortality) trends*
OVER	TIME	7	67	*over time:* e.g. CHANGE/*trends* **** *over time*
AFRICA	SOUTH	7	48	*South (-west/Western) Africa*
ADULT	PREVALENCE	7	43	*adult (HIV) prevalence*
BIRTH	FIRST	7	41	*first birth,* e.g. *age at first birth*
NATIONS	UNITED	7	38	*United Nations*
DHS	SURVEYS	7	33	*DHS surveys, Demographic and Health Surveys (DHS)*
AIDS	CASES	7	30	*AIDS cases*
COUNTRIES	DEVELOPING	7	20	*developing countries*
DURING	PERIOD	7	16	*during* (e.g. *the survey/projection period)*
AFFECTED	BY	7	15	*(heavily/most/also) affected by*
PER	YEAR	7	8	*(infections, per cent, children, deaths* etc.*)*

				per year
FACTORS	SOCIOECONOMIC	7	6	*socioeconomic factors*
MOST	RECENT	7	6	*(the) most recent*
FERTILITY	MARITAL	7	4	*marital (and premarital) fertility*
DATA	RELIABLE	7	3	*reliable data*
MUCH	NOT	7	1	*not much,* C1: *do not differ much from each other*
RESULT	WOULD	7	1	*(this/*NP*) would result in (*NP*)*
AT	RISK	6	100	*at risk of infection/disease, sub-populations at (high/higher)risk*
AGED	WOMEN	6	69	*women aged (15 - 49, 15-24)*
BIRTHS	PREMARITAL	6	32	*(proportion of) premarital births*
LIKELY	MORE	6	28	BE *more likely to* [used to compare different groups of people and their properties]
AFRICA	SOUTHERN	6	24	*Southern Africa*
AGE	SPECIFIC	6	23	*age-specific (mortality rates, fertility rates, prevalence)*
COUNTRIES	OTHER	6	21	*other* (e.g. *African, developing) countries*
LESS	THAN	6	16	*less than*
INFORMATION	PROVIDE	6	15	PROVIDE/INFORMATION
LIFE	TABLES	6	13	*life tables*
BY	PROVIDED	6	9	*provided by*
CHILD	MORTALITY	6	9	*child mortality (*C2 also: *childhood mortality)*
FOR	REASON	6	3	*reason for;* C1: *the/one reason for (this), for any/this/that/some reason,* C2: *the (likely/main)reason for*
BEHAVIOURAL	CHANGE	6	2	*behavioural change*
UNTIL	YEAR	6	1	*until (the) year*

DEMOGRAPHIC	HEALTH	5	22	*demographic and health surveys*
ETHNO-LINGUISTIC	GROUPS	5	21	*ethno-linguistic groups*
HEALTH	SERVICES	5	20	*health services, Ministry of Health and Social Services*
INFECTED	WOMEN	5	18	*HIV-infected women, women (*BE*) infected with HIV*
FOCUS	ON	5	17	*focus on* (both as a noun and a verb)
DEPENDING	ON	5	16	*depending on*
BIRTH	LIFE	5	16	*life expectancy at birth*
CARRIED	OUT	5	12	*(studies, research, census, survey, projection, exercise, campaign, operation) carried out*
PEOPLE	WHO	5	12	*people who*
AIDS	ORPHANS	5	11	*AIDS orphans, orphans as a result of AIDS*
BOUNDS	PLAUSIBILITY	5	10	*plausibility bounds*
EDUCATION	LEVEL	5	8	*level of education*
RATIO	SEX	5	8	*sex ratio*
DIFFER	FROM	5	6	C1:*(not) differ (somewhat, (quite) much) from (each other)*
FOR	INSTANCE	5	6	*for instance*
COMPONENT	METHOD	5	6	*cohort-component method*
FOUND	WAS	5	5	*was found*

CONDUCTED	WAS	5	5	C1&C2: *survey was conducted;*
				C1*: census was conducted;*
				C2*: training, surveillance (round) was conducted*

INFANT	MORTALITY	5	4	*infant (and child) mortality*
ABOUT	ASKED	5	3	*asked about*

FUTURE	NEAR	5	3	*in the near future*
HAVE	IMPROVED	5	3	*have improved* [but valence is different: Maisa uses *improve* without an object, in C2 it is always used with an object]
INFORMATION	RECENT	5	3	C1: PROVIDE *recent information (on/about),* C2: PROVIDE/COLLECT *(recent) information about/on recent*
CONFIDENCE	INTERVALS	5	3	*confidence intervals*
MAKE	ORDER	5	2	*in order to make (*C1: *projection(s), assumptions)*
ASKED	WERE	5	1	(some kind of respondents) *were asked*
MALES	YEARS	5	1	*X years for males and Y for females*
MORE	RAPIDLY	5	1	‘changing’ *more rapidly*
INTERNATIONAL	MIGRATION	4	22	*international migration*
ABLE	TO	4	12	BE *able to*
SPECIFIC	RATES	4	4	*age-specific fertility/mortality rates*
A	REPRESENTING	4	0	*a blue/black curve representing the median/mean*
HAND	ON	3	10	*on the other hand,*
EAST	NORTH	3	7	*North-East*
ACCOUNT	INTO	16	16	TAKE *into account/consideration*
ACCOUNT	TAKEN	11	6	
ASSUMED	IT	21	3	*it* BE*/can be assumed (that)*
ASSUMED	THAT	20	3	
AGE	MEDIAN	11	21	*(increasing/mean/average/high/low/median) age at ((first) marriage, intercourse, birth)*
AGE	MARRIAGE	10	45	
AT	MARRIAGE	8	44	
AGE	MEAN	6	15	

AT	MEAN	5	13	
MEAN	MARRIAGE	5	8	
INFECTIONS	NEW	18	51	*(number of) new (HIV/new HIV/infant) infections*
INFECTIONS	NUMBER	10	19	
FOR	MALES	14	8	(e.g. *life expectancy, adult mortality, rate, ratio) for (males and/than) females*
FOR	FEMALES	12	6	
FEMALES	MALES	9	4	
LIVING	WITH	6	37	*(number/percentage/amount of) people living with HIV*
LIVING	PEOPLE	5	31	
RURAL	URBAN	41	74	*urban and rural* *rural and urban* *urban/rural* *(for) (urban and) rural (populations) separately*
POPULATIONS	RURAL	9	5	
RURAL	SEPARATELY	9	2	
POPULATIONS	URBAN	9	1	
SEPARATELY	URBAN	7	2	
URBAN/RURAL	RURAL	5	13	
AGE	GROUPS	19	62	*age group(s)*
AGE	GROUP	8	58	
PROJECTION	PACKAGE	5	17	*Estimation and Projection Package (EPP)*
ESTIMATION	PACKAGE	5	15	
EPP	PACKAGE	4	16	
ESTIMATION	EPP	4	13	
CHILD	MOTHER	7	26	*mother-to-child transmission*
MOTHER	TO	6	28	
CHILD	TRANSMISSION	6	21	
MOTHER	TRANSMISSION	6	20	
SENTINEL	SURVEY	14	32	*sentinel surveys(s)*
SENTINEL	SURVEYS	13	5	

Table B.2 Maisa, non-matching concgrams: Content-related patterns

Co-occurring word	Co-occurring word	C1	C2	Pattern represented by the concgram
CHILDBEARING	TEENAGE	9	0	*teenage childbearing*
DOUBLING	TIME	8	0	*doubling time*
HIGH	VARIANT	4	0	*high/medium/low variant*
LOW	VARIANT	5	0	
MEDIUM	VARIANT	7	0	
LABOUR	MIGRATION	6	0	*labour migration*
PLACE	RESIDENCE	5	0	*place of (usual/current/childhood) residence*
AN	ESTIMATION	6	3	*an estimation* [part of her title]
DEPENDENCY	RATIO	8	0	*dependency ratio*
ANNUAL	GROWTH	4	0	*annual (population) growth rate*

Table B.3 Maisa, non-matching concgrams: Content-related 'Scenario' pattern

Co-occurring word	Co-occurring word	C1	C2	Pattern represented by the concgram
IF	WERE	13	3	e.g. *if there were* (second conditional)
ANTIRETROVIRAL	IF	5	0	*if there were no antiretroviral treatment* (+ variations)
ANTIRETROVIRAL	THERE	4	0	
IF	TREATMENT	6	0	
IF	THERE	12	9	*if there* BE *(*e.g. *were no AIDS/antiretroviral treatment)*
IF	NO	14	4	*if there* BE *no*
NO	TREATMENT	6	2	*no (antiretroviral treatment)*
ART	NO	10	1	*no ART (scenario)*
NO	SCENARIO	12	1	*No HIV/ART scenario* (+ some variations)
CONSTANT	HIV	32	0	*constant HIV (and declining HIV) scenario(s)*

CONSTANT	SCENARIO	15	0	
CONSTANT	DECLINING	11	0	
CONSTANT	SCENARIOS	10	0	
DECLINING	SCENARIOS	11	0	
DIFFERENT	SCENARIOS	12	0	*different (*) scenarios*
WITHOUT	WOULD	7	0	e.g. *without AIDS,* X *would*

Table B.4: Maisa, non-matching concgrams: Genre-specific patterns

Co-occurring word	Co-occurring word	C1	C2	Pattern represented by the concgram
MY	PROJECTION	9	0	*my projection*
PREVIOUS	STUDIES	5	0	*previous (*) studies*
PREVIOUS	PROJECTIONS	7	0	*previous projections*
I	WILL	11	0	*I will use/introduce/describe/summarize*
I	USE	5	0	*I will use*

Table B.5: Maisa, non-matching concgrams: Individual preferences

Co-occurring word	Co-occurring word	C1	C2	Pattern represented by the concgram
CENT	PER	22	0	C1: *per cent* (BE); C2: *percent* (AmE)
THIS	WAY	12	5	C1: *This way* -[10 out of 12 starting the sentence]; C2: *in this way, way of achieving this, way to do this*
SIZE	STRUCTURE	10	0	C1: *population size and structure/size and structure of the population* C2: *population size, population structure,* but *size and structure* do not co-occur
CAN	DETECTED	9	0	C1: *(impact(s), effect, consequences) can be detected*
CONTINUE	WILL	6	8	*will continue*
EACH	OTHER	6	0	*match/differ/compatible from/with each other*

DHS	STUDIES	6	0	C1: *DHS studies;* C2*: DHS survey(s)*
PAST	YEARS	5	8	C1: *past years;* C2*: in the/over the past* NUMBER *years*/[in tables] *past* NUMBER *years*
BASED	FIGURES	5	0	*figures are based, statistically based figures, based on figures*
ALIVE	MORE	5	0	*more (people) * stay* alive* [in C2 *alive* does not occur]
MADE	UNTIL	5	0	*(projection) made until*
MADE	USING	5	0	*projections/assumptions made using x (software/data)*
AMOUNT	DEATHS	5	0	C1*: amount of (AIDS) deaths;* C2: *number of (AIDS) deaths*
MEDICATION	USE	5	0	C1*:* USE *medication;* C2: TAKE/RECEIVE *medication/medicine(s)*
MUCH	OTHER	5	1	*do not (finally) differ (quite) much from each other*
DIFFER	MUCH	5	0	
YEAR-OLDS	INCIDENCE	3	0	C1: *incidence of* X-Y *year-olds;* C2: *prevalence among* X-Y *year-olds*

Table B.6 Kaisa, matching concgrams

Co-occurring word	Co-occurring word	C1	C2	Pattern represented by the concgram
HISTORICAL	LINGUISTICS	66	13	*historical linguistics*
CHANGES	SOUND	46	4	*sound changes*
BASED	ON	43	61	*based on*
COMPUTATIONAL	LINGUISTICS	34	68	*computational linguistics*
DOES	NOT	23	43	*does not*
BEEN	HAVE	21	50	*have been*
BEEN	HAS	20	26	*has been*
IN	PROTO	20	3	*in (a/the) proto-language*
FORMS	SURFACE	19	19	*surface forms*

LEXICAL	SURFACE	19	13	*lexical forms and surface forms, lexical:surface, the lexical and the surface side*
EACH	OTHER	19	9	*each other*
LANGUAGES	RELATED	18	17	*(genetically, closely, distantly, remotely) related languages*
CORRESPONDENCES	SOUND	17	10	*sound correspondences*
FALSE	FRIENDS	16	112	*false friends*
MORE	THAN	16	36	*more * than*
DO	NOT	16	23	*do not*
COGNATE	PAIRS	15	32	*cognate pairs*
PAIRS	WORD	14	21	*word pairs*
FORMS	LEXICAL	14	14	*lexical forms*
SECTION	THIS	14	13	*this section*
AT	LEAST	13	19	*at least*
LANGUAGE	SPOKEN	13	0	*spoken/language,* e.g. *(once-) spoken language*
COMPARATIVE	METHOD	12	8	*comparative method*
ON	SIDE	12	6	*on the (left, left/right-hand, theoretical, lexical, surface) side*
COGNATE	RECOGNITION	12	2	*cognate recognition*
COGNATES	FALSE	11	94	*cognates and/or/ false friends*
NO	THERE	11	15	*there* BE *no*
CLOSELY	RELATED	11	9	*closely related*
GENETIC	RELATIONSHIP	11	0	*genetic relationship*
MEASURES	SIMILARITY	10	36	*(orthographic) similarity measures, measures of similarity*
EACH	PAIR	10	29	*each (cognate/language) pair*

AS	WELL	10	25	*as well*
M	P	10	11	*the p:m rule/pair*
BY	MEANS	10	6	*by no means, by means of*
BY	FOLLOWED	10	5	*followed by*
COGNATE	LISTS	10	2	*cognate lists*
LINGUISTICS	THEORETICAL	10	1	*theoretical linguistics*
REWRITE	RULES	9	10	*rewrite rules*
ANALYSIS	MORPHOLOGICAL	9	7	*morphological analysis*
FOR	INSTANCE	9	5	*for instance*
EVEN	THOUGH	9	4	*even though*
LEXICAL	SIDE	9	3	*lexical side*
SURFACE	SIDE	9	2	*(the) surface side*
APPLICATIONS	NLP	9	1	*NLP applications*
AUTOMATIC	COGNATE	9	1	*automatic cognate (recognition, detection, identification)*
COGNATES	FRIENDS	8	87	*cognates and/or/ false friends*
FINITE	STATE	8	37	*finite-state (transducers, automata, morphology* etc.*)*
ALONG	WITH	8	7	*along with* (in the sense ‘together with’, e.g. *listed along with X*)
BASIC	VOCABULARY	8	5	*basic vocabulary*
SEVERAL	THERE	8	2	*There are/have been several* + pl. noun *(approaches/attempts/ways* etc.*)*
DESCEND	FROM	8	2	*descend from*
CHANGES	SYSTEMATIC	8	1	e.g. *systematic (sound) changes*
FORM	SURFACE	7	8	*surface form*
COMPUTATIONAL	LINGUISTS	7	7	*computational linguists*
CLOSELY	LANGUAGES	7	5	*closely related languages*

FAR	SO	7	3	*so far*
BETWEEN	RELATIONSHIP	7	3	*(the) relationship between*
FORM	MEANING	7	3	*form-meaning, form+meaning, form and meaning*
LINGUISTIC	THEORIES	7	3	*linguistic theories*
FORMS	MAPPING	7	3	*mapping lexical forms, mapping of forms, mapping from lexical forms*
ALTAIC	URALIC	7	2	C1: *Uralic or Altaic (etymology/languages);* C2*: Uralic and Altaic studies.* [Note that positionally the 'phrase' is fixed.]
FAMILY	LANGUAGE	7	2	*language family*
DONE	HAS	7	1	*has been*
CASE	STUDY	7	1	*case study*
OPEN	SOURCE	7	1	*open source (compiler/program/tools)*
EACH	ENTRY	7	1	*each (word/cognate/vocabulary) entry*
ORTHOGRAPHIC	SIMILARITY	6	36	*orthographic similarity (measures)*
RATHER	THAN	6	24	*rather than*
IT	IMPORTANT	6	10	*it is important to* + inf.
CHARACTER	LEVEL	6	8	*at the character level*
AT	END	6	5	*at the end (of* X)
METHODS	USE	6	4	*methods* USE, *use of computational/statistical methods*
ARE	PERCEIVED	6	3	*are perceived as similar*
M	RULE	6	3	C1: *p:m rule, N:m rule, rule N -> m, rule p -> m;* C2: *p:m rule, N:m rule*
INTO	SPLIT	6	2	C1: *split into two;* C2*: split into*
HARD	IT	6	1	*it is hard to* + inf.
CONCENTRATE	ON	6	1	*concentrate on*

METHODS	USED	6	1	*method(s)/*USE
COGNATE	IDENTIFICATION	5	26	*cognate identification*
DATA	TRAINING	5	19	*the training data*
ACCOUNT	INTO	5	10	TAKE *into account*
LENGTH	WORD	5	10	*word length, the length of a/the word*
COMMONLY	USED	5	9	*commonly used*
LATIN	VULGAR	5	5	*Vulgar Latin*
BY	CAUSED	5	4	*caused by*
AT	STAGE	5	4	*at (the, this, each, the next, the final) stage*
MAKES	POSSIBLE	5	3	NP*/this make it possible to* + inf.
MANY	THERE	5	3	*There* BE *(not) many* +NP
HAVE	UNDERGONE	5	3	*have undergone/change (or shift)*
BEEN	DONE	5	2	*been done*
BETWEEN	RELATION	5	2	*relation between (languages)*
MEANS	NO	5	1	*by no means*
AT	HAND	5	1	NP+ *at hand*
FAMILY	TREE	5	1	*family tree*
FRONT	VOWELS	5	1	*front vowels*
RULES	REFER	5	0	*rule(s) + to refer* [C2: 1 instance]
APPROACHES	COMPUTATIONAL	5	0	*computational approaches (to)*
AT	LEVEL	4	15	*at the (feature, character, sentence, proto-language) level*
CONCLUDE	THAT	4	6	C1: *the writers/authors conclude that;* C2: *we conclude that (ant other uses)*
LINGUISTIC	THEORY	4	2	*linguistic theory*
BACK	VOWELS	4	1	*back vowels*

LENGTH	VOWEL	4	1	*vowel length/length of the vowel*
AND	PAPER	24	18	*paper-and-pencil (linguistics)*
AND	PENCIL	24	13	
PENCIL	PAPER	16	11	
PAPER-AND	AND	14	11	
PAPER-AND	PENCIL	14	11	
AND-PENCIL	PAPER	14	10	
LINGUISTICS	PAPER	6	5	
PAPER-AND	LINGUISTICS	6	5	
PENCIL	LINGUISTICS	6	5	
AND-PENCIL	LINGUISTICS	6	4	
PAPER-AND-PENCIL	LINGUISTICS	6	4	
PENCIL	LINGUISTS	8	3	*paper-and-pencil linguists*
AND-PENCIL	LINGUISTS	8	3	
LINGUISTS	PAPER	8	2	
PAPER-AND-PENCIL	LINGUISTS	6	3	
PAPER-AND	LINGUISTS	6	2	
LANGUAGE	PROTO	63	2	*proto-language*
PROTO-LANGUAGE	LANGUAGE	56	1	
LEVEL	TWO	65	47	*two-level (rules, model, morphology, compiler, grammars)*
LEVEL	MODEL	11	8	*two-level model*
MODEL	TWO	11	6	
TWO-LEVEL	MODEL	10	5	
MORPHOLOGY	TWO	6	11	*two-level morphology*
LEVEL	MORPHOLOGY	6	11	

TWO-LEVEL	MORPHOLOGY	5	10	
RULES	TWO	40	18	*two-level rule(s)*
LEVEL	RULES	38	15	
TWO-LEVEL	RULES	35	15	
LEVEL	RULE	6	4	
COMPILER	TWO	5	5	*two-level (rule) compiler, compiler for two-level rules*
LEVEL	COMPILER	5	5	

Table B.7 Kaisa, non-matching concgrams: Content-related patterns

Co-occurring word	Co-occurring word	C1	C2	Pattern represented by the concgram
COMPUTATIONAL	HISTORICAL	29	0	*computational historical linguistics, computational methods in historical linguistics, computational and historical linguistics*
PROTO	WORD	20	2	*proto-word*
PROTO	SOUND	20	0	*proto-sound*
UGRIC	FINNO	16	0	*(proto-)Finno-Ugric*
COMPUTATIONAL	METHODS	10	0	*computational methods (in historical linguistics)*
FINNIC	SUMERIAN	10	0	*Sumerian-Finnic/Sumerian and Finnic (entry/ies)*
RULE	VARIABLES	10	0	*rule-variables*
LANGUAGE	ONCE	9	0	*(once-) spoken/language*
DAUGHTER	LANGUAGES	8	0	*daughter languages*
FINNISH	UDMURT	8	0	*Finnish and Udmurt*
LANGUAGE	LIVING	7	0	*a living language* [Kaisa writes about Sumerian]
PROTO	WORDS	7	0	C1: *proto-words;* C2: *proto-language(s), -projections, -phonemes*

ALTAIC	OR	7	0	*Uralic or Altaic (etymology)*
ETYMOLOGIES	URALIC	6	0	*Uralic etymology(ies)*
FINNO-UGRIC	PROTO	6	0	*Proto-Finno-Ugric*
GRAMMARS	TWO	6	0	*two-level grammars*
ONCE	SPOKEN	6	0	*once-spoken language, once a real/living spoken language*
UGRIC	PROTO	6	0	*Proto-Finno-Ugric*
SOUNDS	PROTO	6	0	*proto-sounds*
ANY	URALIC	6	0	*any (known/attested form of a) Uralic language*
DATA	ORIGINAL	6	0	*original data*
SHORTER	THAN	5	2	*shorter than*
ENVIRONMENTS	PHONETIC	5	0	*phonetic environments*
FIELDS	LINGUISTICS	5	0	*(the two) fields of linguistics*
INVENTORIES	PHONEMIC	5	0	*phonemic inventories*
FORM	PROTO	5	0	*proto(-language)-form*
ANY	LIVING	5	0	*any living language, any language living or dead, any * living affiliates/relatives*
FILE	GRAMMAR	5	0	*grammar file*
ETYMOLOGY	WORDS	4	0	*word with* X *etymolog(y/ies)*
CALLED	RULES	4	0	*are called* X *rules*
ALPHABET	I	4	0	*in the Alphabet* (11/0)
LIST	VALUE	3	0	*value list*
LONG	VOWELS	3	0	*long vowels*
PROTO	RECONSTRUCT	6	0	C1: RECONSTRUCT/*proto-sound(language);* C2: RECONSTRUCT/*history, languages,* (*proto-phonemes* – 1 instance)
PROTO	RECONSTRUCTED	8	0	e.g. *reconstructed proto-language/sound* (*proto-language can be reconstructed*)

LANGUAGE	RECONSTRUCTING	6	1	*reconstructing a proto-language*
PROTO	RECONSTRUCTING	9	1	
PROTO-LANGUAGE	RECONSTRUCTING	5	0	
RECONSTRUCTED	SOUND	8	0	*(the) reconstructed sound*

Table B.8 Kaisa, non-matching concgrams: Genre-specific patterns

Co-occurring word	Co-occurring word	C1	C2	Pattern represented by the concgram
I	WILL	27	2	*I will*
FIRST	WILL	7	2	*first I will, I will first*
SECTION	WILL	5	1	*section will/ In this section I will*
CONCENTRATE	WILL	5	0	*I will concentrate*

Table B.9 Kaisa, non-matching concgrams: Individual preferences

Co-occurring word	Co-occurring word	C1	C2	Pattern represented by the concgram
CORRESPONDENCE	SOUND	13	0	*sound correspondence (set(s)*
GENETIC	RELATIONSHIP	11	0	*genetic relationship*
ANALYSIS	GENERATION	10	0	*analysis and generation* (5 instances)
FED	INTO	8	0	*data/cognates/list of words and morphemes/material fed into (system/entry/database)*
POSTULATED	SOUND	8	0	*postulated proto-sounds/sound changes*
AUTOMATIC	RECOGNITION	7	0	*automatic cognate recognition*
GO	THROUGH	6	0	*I will go through X*
BE	NEEDS	6	0	*needs to be* +verb-*ed*
SYSTEMATIC	WAY	6	0	*systematic way, in a systematic way*
ABOVE	EXAMPLE	6	0	*the example above*

MORPHEMES	WORDS	6	0	*words and morphemes*
AID	NLP	5	0	*aid NLP applications*
DISTANT	RELATIONSHIP	5	0	*distant (genetic) relationship*
FROM	MODIFIED	5	0	*modified from* + source [used in 3 different drafts]
CHANGES	POSTULATED	5	0	*postulate/sound changes*
APPLICATIONS	USED	5	0	*something used to (help) aid/refine NLP applications*
NLP	USED	5	0	*X used to aid/refine NLP applications*
STILL	WAS	4	2	*was still*
LEVEL	USE	4	1	*use of two-level morphology/rules*
AFFILIATION	LINGUISTIC	4	0	*linguistic affiliation*
AND	MEDIAL	8	0	*(word-) initial, medial and final (sound) changes*
CHANGES	INITIAL	6	0	
INITIAL	WORD	5	1	
INITIAL	MEDIAL	5	0	
E	FRONT	6	0	*(Y * realised as) i * in front of an epenthetic e*
AN	FRONT	5	0	
EPENTHETIC	I	5	0	
EPENTHETIC	FRONT	5	0	
FRONT	I	5	0	
FRONT	Y	5	0	
E	EPENTHETIC	12	0	*an epenthetic e*
AN	E	9	3	
AN	EPENTHETIC	9	0	
HISTORICAL	KNOWLEDGE	5	0	C1*: knowledge of historical linguistics; C2: knowledge of X*
KNOWLEDGE	LINGUISTICS	5	0	

COGNATES	PROTO	8	1	*mapping/building/generating (from) cognates/proto-word from/to cognates/proto-word*
MAPPING	PROTO	7	0	
FOLLOWING	WILL	12	0	*In the following I will + verb*
FOLLOWING	I	11	0	
COMPUTATIONAL	PAPER	6	2	*computational and paper-and-pencil-linguistics/linguists*
COMPUTATIONAL	PENCIL	6	2	
AS	REALIZED	8	0	X *(*BE/GET*) realized as* Y
AS	REALISED	5	0	

Appendix C: Meaning-based responses are ‘harder’ to give

Table C.1 Stimulus-response pairs which were ‘hard’ for the respondents

Stimulus-response	Comment	M/ S?	How does it match the usage?	MWU(s) in usage
Maisa				
rural – countryside	think of countryside, I was thinking about this article I was writing about in here yesterday but it was so hard to have a word, I was thinking about some figures probably...	M	Non-matching	*rural populations* *rural and urban (areas)*
Nora				
important – for me	that was I think it was hard, so that was just important for me	S	Matching, colligation	*important for smb/smth*
lays – somewhere	that was also, it’s just nothing strictly came to my mind	S	Matching, colligation	*(the) main focus* LAY *on* + NP (4 out of 5 occurrences of LAY)
accepted – o.k.	that was also hard, because I just went through and wrote that one and there were some word there was nothing immediately coming and that was one of them, so it’s just like yeah accepted – it’s ok	M	No MWU	
flexibility – not fixed	that was also a bit hard, after a moment of stopping I put not fixed	M	No MWU	
underlies – lies under	that was also hard for some reason, so I just lies under	M	No MWU	
hereby – with this	it is also something I don’t use it often but I know it in context but it was blank for a moment in my mind	M	No MWU	
Linda				
consistent – sure	I guess, you are sure of something if you are consistent, that was a hard one	M	Non-matching	*consistent with*
situation – place	a place, I don’t really know, that was a hard one too, situation.. yeah it is a place if there is a situation	M	Non-matching, core meaning	*situation analysis*
there – here	it was kind of hard, I don’t know why, it’s like ok, you are there, so I am here	M	Non-matching, core	*There is no need to* + inf.

			meaning	
somewhat – certain	that was hard, I guess I thought certain would be the opposite, somewhat there, somewhat here , when you are certain it is not somewhat anymore	M	No MWU	
whereas – when	that was a hard one too, these two were hard because I don't really know any synonyms – somewhat and whereas – yes.. so it is hard to even explain why is when,	M	No MWU	
Hertta				
context - burial	that was surprisingly difficult because I talk about context all the time and of some reason it became stop and I just wrote burial because it is a context quite closed context often	S	Matching, YX	*burial contexts*
trial – excavation	was not that easy, I paired with excavation, I know I use a trial excavation and before I have actually.. that was a word that a better supervisor suggested because I had written test and she marked that it should be a trial and after that I have sort of noticed that when I read texts that it is a trial excavation Funny.	S	Matching	*trial excavation(s)*
adults – male	that was not easy, so I just wrote male	M	No MWU	
child – infant	that was quite hard. Something that goes with child, but then it became infant because it is sort of at least in the same age group	M	No MWU	
vary - difference	and that one I was quite unsure what to write because I was thinking another word but then I remembered that this was the first word so that why it looks like, the difference is difficult not sort of that fluent and I really don't know why I thought about difference, and now I actually can't remember what was the second word that made	M	No MWU	

	more sense to vary			
possibly - possibility	that wasn't the easiest word either. that comes out as a if I have possibly a possibility or something	M	No MWU	
located - on	I was not quite sure what to write so it just became on, something might be located on something	S	Matching, colligation	BE *located at/behind/in/on*
Kaisa				
so - that way WAT5	so is also a hard one and I wrote that way, sort of synonyms	M	Non-matching	*so far, so that*
otherwise – or WAT5	that is also a hard one this otherwise, that is why... I mean or does not really fit	M	Non-matching	a colligation with a negative
pinpoint - point out WAT5	that was also pretty hard because point out is a bit more vague than pinpoint	M	Non-matching	colligation with an object and a semantic preference for some kind of 'infelicities'
beyond –behind WAT5	that was also pretty hard, beyond this ... then it is not really behind this, [so actually the association was beyond THIS?] Exactly	M	Non-matching	GO/SCOPE/EXPAND *way/far beyond* NP
would – should WAT5	hard one to come up with also	M	Non-matching	*would have,* WOULD/IF
along – by WAT6	that was a harder one	M	Non-matching	*the lines of, along with* + NP meaning 'together with'
have – hold WAT6	that was a bit hard but it is like if you have something like physically have then it can sometimes mean that you hold	M	Non-matching	*have been, have undergone, would have*
and – or WAT5	that is also hard to come up with	M	No MWU	
moreover – still WAT5, N103	that was also hard this moreover, cause it does not really associate with anything	M	No MWU	
however - even though WAT5, N112	hard, such words are hard to come up with... because you use them so consistently..or you usually don't think of any variation to that [-so they kind of don't associate with anything] No [-stand on their own] yes. not even with	M	No MWU	

	any phrases, it is usually like “however” comma, there is nothing that goes along with them			
thus - that’s why WAT5	that is also hard this thus	M	No MWU	
yet – still WAT5	that was also pretty hard, but not as hard as the other ones, sort of use it as a synonym to still. Yet I have this to still uncover, Still I have this to yet uncover. No. well sort of	M	No MWU	
whereas – but WAT5	that is also a hard one... because there is ...no, I don’t know why I wrote but, it does not really fit	M	No MWU	
whereas – but WAT6	a bit hard, it is not really fitty	M	No MWU	
also – plus WAT6	that was hard	M	No MWU	
since – because WAT6	hard one	M	No MWU	
specifically – precisely WAT6	that was a bit hard,[because it is an adverb?] adverbs are pretty hard or I don’t know, specific , it is such a specific word	M	No MWU	