CHAPTER 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 3,600 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]: 137–8). This casts doubt on the plausibility of independent lexical choice and suggests that words are not 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 per cent 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 are not the numbers of words one can feasibly acquire through explicit instruction and studying. Most learning must happen through exposure, i.e. most learning is likely to be usage-based.

These are some of the reasons for a vast and still growing (see e.g. Christiansen and Arnon 2017) interest in procedural knowledge of units larger than a word acquired (implicitly) through exposure both in first and second language processing and use. This interest spans many (if not all) fields in linguistics and cognitive sciences ranging from corpus linguistics and psycholinguistics to applied linguistics and second language acquisition. Given this state of play, the topic can clearly benefit from interdisciplinary cross-fertilisation. What I will attempt to do in this book is take a descriptive linguistic approach but build it as much as possible on current knowledge from a variety of often separately developing disciplines, such as, corpus linguistics, psycholinguistics, cognitive science, first and second language acquisition (SLA), English as a lingua franca (ELF), language variation and change.

There are numerous descriptions of units larger than a word or multiword units (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 word 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

[1] www.oxforddictionaries.com

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: 20). Indeed, by declaring '[t]he phrase, the whole phrase and nothing but the phrase' (Sinclair 2008: 407), Sinclair puts forward phraseological patterning as a characteristic property of language as a whole.

In fact, the focus shifts from entities, items which have a 'special status', to the process, the cognitive mechanism of language production by co-selection, known as the idiom principle (Sinclair 1987). There are a number of other similar concepts which were proposed in different fields of linguistics and psycholinguistics. In principle it is likely to be the same process which is referred to as chunking, holistic processing, Hebbian or associative learning and sequential processing, as the following quotes from a few authors illustrate:

> Chunking is the process by which sequences of units that are used together cohere to form more complex units. (Bybee 2010: 7)

> Speakers do at least as much remembering as they do putting together. (Bolinger 1976: 2)

> The general idea is an old one, that any two cells or systems of cells that are repeatedly active at the same time will tend to become 'associated', so that activity in one facilitates activity in the other. (Hebb 1949: 70)

> [Formulaic sequence is] 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)

> Sequential processing refers to the mental process of integrating and understanding stimuli in a particular, serial order. Both the perception of stimuli in sequence and the subsequent production of information in a specific arrangement fall under successive processing. (Huang 2011: 2261–2)

Thus, it is hoped that while the book describes the process underlying phraseological patterning in terms of the idiom principle, readers who come from other paradigms of linguistic theory will also find the book relevant to their interests.

One of the questions which varying terminology raises is how closely we can actually describe what is happening in language processing. For example, in a recent article, Siyanova-Chanturia (2015) criticises the application of the term 'holistic processing' arguing that the fact that formulaic units are processed faster than non-formulaic does not yet mean that they are

processed 'unanalysed' or 'holistically'.[2] While indeed it is very important to stay critical to the terminology we are using and bear in mind the implicit assumptions it can generate, very often we do not have much more than just some variant of metaphoric language at our disposal. So while I agree that terms like holistic processing, mental lexicon, representations in the mind, storage and the like do not capture what actually happens in the brain, I will have to resort to these terms now and then for lack of better alternatives.

In fact, I agree with Casasanto and Lupyan (2015) who argue that all concepts are ad hoc concepts, or in other words language processing and use can only be described in terms of processes rather than stored representations:[3]

> Rather than trying to explain concepts, categories, and word meanings as things that we *have* in our minds, like entries in a mental dictionary or mental encyclopedia, it may be more fruitful to build theories of conceptualizing, categorizing, and constructing word meanings: things that we *do with* our minds. (Casasanto and Lupyan 2015: 543; emphasis in the original)

From a certain perspective, Sinclair's position is similar: he argues that the lexicon is actually 'empty' emphasising 'variability of the units of the "live" lexicon [... which] adapt to the ever-changing, never-quite-repeated circumstances of communication, and as such cannot, in principle, be fully prescribed in advance' (Sinclair 1996b: 161; see also Elman 2009). Thus, instead of producing a list of MWUs or focusing on their identification, he develops a model of a unit of meaning which allows for inherent variability and describes smaller processes which accompany co-selection.

Sinclair predicts that a unit which is produced on the idiom principle will be categorised by one meaning, a set of specific properties and, importantly, it will undergo meaning shift. Thus, while he first described the model of a unit produced on the idiom principle as a unit of meaning (Sinclair 1996a), he later suggested a new term – a meaning-shift unit (see Cheng et al. 2009). In this book, I will use both terms but will give pride of

[2] In my approach, in some way similar to Arnon and Christiansen (2017), an individual's processing a MWU on the idiom principle does not preclude his/her access to separate words forming it. At the same time, the likelihood of the individual 'noticing' the separate words within a MWU is conditioned by the degree of delexicalisation of the unit (see Sections 2.3 and 2.4). One and the same MWU can be read both on the idiom and on the open-choice principle (see Section 7.3).

[3] See also Ellis (2008) who 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).

place to the meaning-shift unit (MSU) as the inherent potential for variation and change and the dynamic nature of the unit, which will be important for us here, is made explicit in this new term. The possibility for abbreviation 'MSU' proves handy too. The intrinsic connection between the idiom principle and meaning shift will be discussed in detail in Chapter 2.

In some way, meaning-shift units also have a link to larger cognitive phenomena of prediction (see e.g. Clark 2013) and statistical learning (see e.g. Rebuschat and Williams 2012) in human (language) processing. However, the type of regularities captured by the model is quite clearly not the only type of regularities manifest in language, and similarly it is not the only type of regularities on which prediction can be based and which can be learned implicitly through exposure. Regularities are present at different levels of abstraction and scale. There is also no single 'best' description of a multi-word unit: different descriptions capture specific regularities, but the human brain is likely to be sensitive to all of them, to a different degree and with variation across individuals. The meaning-shift unit is a type of regularity which is likely to play a role in prediction and statistical learning.

Cognitive linguistic literature makes an important distinction between explicit and implicit learning which happens subconsciously without attentional control (or even without the learner noticing the learning) based on inferences from regularities present in the input. However, in processing natural language, the mind probably switches back and forth between implicit and explicit learning all the time. In this book I simply focus on language use and learning from exposure in a natural language environment. Such learning is likely to involve implicit learning but is not equal to it. Therefore the results of this study with respect to implicit learning need to be interpreted with caution.[4]

While there is no question that phraseological patterning is likely to be key both in language organisation and in language processing, second language (L2) learners are usually excluded from this picture as an exceptional case. The attention of SLA scholars is captured by the problems L2 learners 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 in grammar and knowledgeable

[4] Cf. also the notions of incidental and intentional learning used most widely in vocabulary studies (e.g. Hulstijn 2012).

at the level of words, would always be a potential victim of that lesser store of formulaic sequences' (Wray 2002: 210). The ability to acquire, process and use MWUs holistically appears to be compromised in the case of L2 learning. In other words, 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; Wray 2002; Seidlhofer 2009). This apparent paradox serves as the background for the study presented in this book.

Thus, in this book, my aim will be to explore meaning-shift units (MSUs) in second language usage, acquisition and processing. Importantly, I will argue for the value of looking at individual languages and tracing MSUs as they are learned from exposure, used in individual language output and processed in the mind intra- rather than inter-individually. After all, we are interested in the cognitive mechanism of the idiom principle which is a property of an individual mind rather than the collective average (see a detailed discussion of this argument in Chapter 4). In the study presented in the book, three types of data were collected from five non-native female 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. This research design enabled qualitative comparison of lexical usage patterns, conceived of in accordance with John Sinclair's conceptualisation of lexis and meaning, to (1) language exposure and (2) word association responses.

The reader might legitimately question to what extent a multilingual student of a highly international university writing a Master's in English is a language learner, the term used in the title of the book. Indeed this term underlines the process of learning and so implies that the person is still in that process, in contrast to, say, a native speaker who has allegedly 'completed' the process. The implication puts a language learner in an inferior position, discredits her language proficiency and denies the status of an independent language user. Yet, a native speaker is not always highly proficient nor an L2 learner always a school-aged child with a textbook in the classroom. Such controversy gives rise to a strong argument in favour of the term 'L2 user' (Cook 1999, 2002) or even the terms 'L1 user' and 'LX user' and doing away with the term 'native speaker' altogether (Dewaele 2018). In addition to this, as Mauranen (2011, 2012) points out, pedagogical

contexts, like the classroom, set very different targets for language learners, making them 'display' their language knowledge and orient towards accuracy and Standard English norms. Outside the classroom and especially in English as a lingua franca communication such norms are often of little relevance and instead mutual understanding becomes the main goal (see also the discussion in Section 3.3). As such, language learners and language users may end up using language differently. In this sense, the participants of this study are language users rather than language learners: they did not write theses to practise their language skills and in fact throughout the book they are referred to as language users. Still it was deemed necessary to leave the term 'language learner' in the title for two main reasons. The first reason has already been mentioned: the term underlines the process of learning a language and this is the focus of the book: not in the sense of attaining competence but in the sense that using language and learning language are inseparable, for L1 learners and LX learners alike.[5] Secondly, the book engages in dialogue with previous studies on the idiom principle in second language acquisition and use of MWUs, all of which, except for a small fraction of those on ELF use, are based on the dichotomy between native speakers and language learners (or non-native speakers) and use the terms accordingly. With the aim of building a constructive dialogue, the terminology is preserved.

The book is organised as follows. Chapter 2 focuses on the main concept of this book, the meaning-shift unit, and offers an in-depth discussion of Sinclair's concepts: the idiom principle, collocation, colligation, semantic preference, unit of meaning, meaning-shift unit, semantic prosody, ultimate dictionary, core meaning and delexicalisation and shows how they form a coherent framework of conceptualising lexis and meaning. This exercise forms the theoretical backbone of the book. Chapter 3 then moves on to review studies on L2 use and processing of MWUs. It transpires that there is both evidence supporting the 'collocational deficiency' hypothesis and conflicting with it. A different line of argument can be found in English as a lingua franca research which underscores variability as one of the main features of L2 use and offers a different explanation of it. Chapter 4 presents the three types of data that will be analysed in this study including the theoretical motivation behind each of them and argues for the value of intra-individual comparisons. The empirical work in this study is divided into two parts. First, usage patterns are compared to exposure in Chapter 5. Then, usage patterns are compared to word association

[5] See e.g. Dąbrowska (forthcoming) for the argument about the impact of print exposure on language representation, and ultimately competence.

responses in Chapter 6. Following the descriptive aims of the study, both comparisons are predominately qualitative and explore the structure and processes working within meaning-shift units. Some simple quantitative comparisons are also provided, especially in Section 5.2 and Section 6.3.6. Chapter 7 summarises the findings first spelled out in the respective chapters and integrates them into the model of a meaning-shift unit with its associated processes.

CHAPTER 2

From a Unit of Meaning to a Meaning-Shift Unit

In the last three decades, multi-word units (MWUs) of all kinds have attracted a wealth of attention from researchers of different subfields of linguistics and psycholinguistics. Explorations of what is often called the phraseological phenomenon have generated an impressive variety of terms: from idioms, chunks, clichés, routines, lexical phrases, fixed expressions, fossilised forms, irregular phrases to phraseological units, multi-word expressions, formulaic sequences, prefabricated patterns and collocations, to name but a few (see also Wray 2002: 9). It is disputable whether these terms indeed focus on the same phenomenon, but what seems to be common to all of them is that they emphasise the special status of some units. As Wray points out in her well-known 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 thus can pinpoint its boundaries, or we calculate that the item as a whole means something different from 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 the availability of 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 organisation in language in use' (Stubbs 2009: 115).

The innovation corpus linguistics has brought with it lies in the new kind of observation it makes possible: in short, it enables us to see 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 provides and the kind of observation that led Darwin to his theory of species. Apparently, a drawing of finches from the Galápagos Islands where they were 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 bird family, while the differences arose as a consequence 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, suggesting that a whole number of word sequences are in fact instances of one and the same pattern (cf. Sinclair 2004 [1998]: 148).

Such 'disturbing' (Sinclair 2004 [1998]: 133) corpus evidence has led John Sinclair, one of the pioneers of corpus linguistics, to develop an entirely new conceptualisation of lexis and meaning, which gave rise to such well-known concepts as the idiom principle and unit of meaning. This chapter offers a step-by-step interpretation of this conceptualisation which is then used as a theoretical framework for the empirical study of meaning-shift units in L2 acquisition, processing and use in Chapters 5 and 6. Special attention is 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.

2.1 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