Corpora, phraseology and dictionaries: How does corpus research intersect language teaching and learning?

JARMO HARRI JANTUNEN
University of Jyväskylä

This article discusses the role of corpus data in language learning and teaching as well as the benefits of using authentic language data in learner dictionary writing. It has been argued that acquiring and teaching a target language and its phraseology would benefit from the usage of naturally occurring language. In the research on learner language phraseology to date, there is a bias towards analyses of collocations and *n*-grams. The present article attempts, however, to widen this scope to other dimensions of phraseology, namely semantic preference and semantic prosody, which are more abstract and perhaps more difficult to learn than concrete co-occurring lexical items. This article introduces a case study of the Finnish degree modifier *oikein* ('very'), which often occurs in deviant phraseological patterns in learner writing. The data come from the International Corpus of Learner Finnish and the Corpus of Translated Finnish. The article suggests how Finnish learner dictionaries could represent words as phraseological items.

Keywords: corpora, language teaching, dictionaries, learner dictionaries, phraseology, collocations, *n*-grams, semantic preference, semantic prosody.

1. Corpus research and second/foreign language learning and teaching

The so-called native language corpora are often compiled and especially analysed in order to reveal contemporary usage of native/target language. Furthermore, contrastive corpus analyses are made between learners' mother tongues and target languages with a view to hypothesising the difficulties that learners may face in language learning or in order to explain observed problems (for an overview of corpus-based contrastive learner language analyses, see Granger 2015). In comparison, learner corpora – i.e. databases which typically include material produced by foreign or second language learners in language learning situations and which "allow learners to choose their own wording rather than being requested to produce a particular word or structure" (Paquot & Granger 2012: 131) – focus on learner outcome and are compiled to reveal patterns that learners produce and difficulties they have faced in communication. However, although a vast number of native and learner corpora exist, they are still relatively seldom utilised in language learning and instruction. This statement may be surprising in light of Römer's (2008: 112) argument, which suggests that corpus linguistics and language pedagogy form a dynamic relationship and that corpora and corpus-based materials are available for learners and teachers. Römer's argument may be true if the focus is on English language learning and teaching, but the situation is often different when discussing less frequently taught and minor languages, such as Finnish.

Possible reasons for the lack of corpus-based methods and materials in language instruction are discussed, for example, by Meunier (2011). She highlights, in particular, four reasons for this lack (461–463):

1. Corpora, corpus methods and corpus studies have not reached teachers, or may be rejected by them.
2. Using authentic data is only one pedagogical method.
3. The importance of frequency, i.e. are frequent items also useful to teach?
4. The lack of studies exploring the impact of corpus methods on language learning.

Corpora are used as a source in reference and pedagogical materials in some languages, but practice in general is still largely unaffected by corpus research. In the discussion regarding the combination of corpus linguistics with language learning and teaching practices, there are several points to consider. First, it is not clear what teachers know about the types of corpora that are available in certain languages. Several web-based listings of corpora exist, but it is also not known if teachers are familiar with them. In addition, teachers may not be aware of particular areal, societal and, for example, time- and genre-based criteria used in compilation processes. Such criteria are usually included in corpus metadata (e.g. source of the data, corpus components, annotation, availability, licences), but it is not self-evident that possible users of corpora are used to seeking out that kind of information. If a suitable corpus is found, tools for corpus analyses (in the form of software packages) must be acquired. Although modern corpus tools are user friendly, some training is usually needed at this stage. Furthermore, skills in corpus methods, that is, certain techniques to analyse the data, are also a requirement, as is, finally, knowledge of how to interpret the information. It seems that a range of skills are needed from teachers as well as from learners if they want to benefit from corpora. It then becomes easy to argue that, in line with Meunier (2011) and Römer (2008), corpora and corpus methods are in less frequent use in the classroom than they could be.

Both Römer (2008) and Meunier (2011) suggest several solutions that would help increase the role of corpora in language instruction and learning. Closer, cross-disciplinary integration between corpus linguists, SLA researchers, and language teachers would benefit from these paradigms and lead to a better understanding of the needs of teachers and learners. This means, for example, direct corpus use in classroom instruction or indirect use in developing dictionary and textbook materials (see e.g. Chambers 2007), as well as corpus use in language teacher education and elsewhere. Another suggestion is to increase the use of triangulation methods. This could mean, for instance, combining corpora and other kinds of data and research methods. As stated above, teachers and learners are often unfamiliar with corpora, a situation which calls for the clear and step-by-step teaching of methods for using them. There is also a need to develop corpora, native as well as learner, in order to "find out more about the characteristics of learner language, so that, in the

future, a larger number of dictionaries, grammars, and textbooks will not only be corpus- but also learner corpus-informed" (Römer 2008: 123). Classroom instruction would also benefit from direct learner corpus use: where native data reveals which items and structures are favoured (or avoided) by target language users, learner data can be used to provide information on problematic, deviant or atypical language usage; the most beneficial method would be to let learners compare these two sets of data, find differences and study language in a data-driven manner. The integration of corpus collection and data analysis with classroom activities would lower the threshold in using data-driven methods in learning and make the usage of corpus data more attractive. Offering qualitative information alongside frequency information helps learners to practice and develop phraseological and grammatical skills, especially because frequency and statistics do not often interest learners and teachers. However, before corpora and corpus methods can find their place in classroom interaction, corpus researchers must analyse corpus data and items and produce information that is relevant for teachers and language learners. This demands, first of all, close collaboration between teachers and corpus linguists.

The remainder of this article is divided into four sections. Section two describes theoretical perspectives on phraseology and multiword units, connecting them to language learning and corpus linguistics. Section three introduces a case study which describes the Finnish degree modifier *oikein* as a phraseological unit in learner data in order to illustrate difficulties that learners face with phraseology. The following section focuses on the role and need of phraseology in compiling learner dictionaries, taking the previous analysis of *oikein* as an example. Finally, section five briefly discusses the conclusions that can be drawn from the case study as well as directions for future dictionary making, language education and research.

2. Corpora, phraseology and language learning

In the mid-1980s, the phraseological view of language and the use of prefabricated language became of interest in language learning studies, mostly in studies focusing on English language learning (see e.g. Granger 1998). After the 1990s, the existence, description and use of prefabs, collocations (i.e. "the occurrence of two or more words within a short space of each other in a text", Sinclair 1991: 170), multiword units and formulas (for terms used to refer to formulaic language, see e.g. Wray & Perkins 2000; Wray 2002) also attracted the interest of teachers and researchers outside of the EFL world. The so-called phraseological boom (Thewissen 2008; Jantunen 2009; see also Paquot & Granger 2012) can be continuously seen in several publications and is gradually spreading in the teaching and study of languages other than English as well.

Several corpus-based studies on learner language (e.g. Granger 1998; Nesselhauf 2005, 2009; Jantunen 2009; for overviews of corpus research in an L2 English context, see Paquot & Granger 2012; Granger & Meunier 2008) have observed that learners face clear problems with the phraseology of the target language: they either overuse or underuse prefabricated patterns available in the target language and produce deviant phraseological combinations. The reasons for this may be that learners partly produce language more in line with Sinclair's so-called open-choice principle than they do following the idiom principle (see Sinclair 1987: 319) or crosslinguistic influence (transfer of form, frequency and register from the learner's L1 or other languages acquired earlier, see Paquot 2008, Nesselhauf 2005). A recent study by Vetchinnikova (2014: 216), however, suggests that learners tend to produce multiword units according to the idiom principle more often than was assumed earlier and that "the phraseological ability of L2 speakers does not seem to be fundamentally different from NS ability". This is contrary to previous claims by, for instance, Kjellmer (1991: 124), who has stated that since learners have automated only a few target language collocations, "their building material is individual bricks rather than prefabricated sections". Nevertheless, multiword units still cause problems to some extent and the uncertainty of using prefabs does not resolve easily, thus meaning that atypical phraseology remains problematic during the whole learning process: collocations, for example, are difficult even for advanced learners despite mastery of grammar and other areas. Nesselhauf's (2005) study also shows that neither time spent in classroom teaching nor exposure to the L2 in the target language society has a clear positive effect on collocational accuracy. However, the analysis of phraseological problems should move beyond the collocational level and take into account the more abstract, and perhaps more challenging, phraseological features. These are semantic associations, such as the semantic preferences (i.e. the co-occurrence of a lexical item with words sharing similar semantic features, the semantic field that collocates belong to) and semantic prosodies (abstract general association of words and their positive or negative cotext, which expresses attitude or evaluation) of lexical items (see e.g. Sinclair 1991, 1996; Stubbs 1995; Steward 2010). These phraseological dimensions have not been thoroughly studied in the learner phraseology research even though they are essential features of lexical items and ought to be taught and mastered along with collocations. The present article endeavours to provide a more holistic analysis of phraseology in learner language, with an additional focus on semantic associations.

In addition, in morphologically rich languages, the mastering of morphological preferences (see Jantunen & Brunni 2013) essentially belongs to the learning of phraseological grammatical dimensions in conjunction with other grammatical dimensions, such as colligations (i.e. the typical

grammatical patterning of words, see e.g. Hoey 1997: 4). Therefore, what is known to date about the phraseological problems that learners face is still fairly imperfect. Unfortunately, phraseological dimensions other than collocations are only seldom discussed in learner phraseology studies (see Wang & Wang 2005; Wei 2006; Xiao & McEnery 2006; Ahmadian et al. 2011). To sum up, despite Granger's statement from 1998 (159) that "prefab-oriented approaches to teaching are currently, in the late 1990s, in vogue", and that phraseological problems in learning are today a well-known challenge, much ought to be done in teaching phraseological dimensions other than collocations and in languages other than English.

3. Case study: *oikein* ('very') in learner Finnish

The following case study of the degree modifier *oikein* ('very') illustrates some of the aspects discussed earlier in this article. The case tries to shed light on the phraseological problems that learners of Finnish have with degree modifiers, but it also endeavours to widen the analysis of phraseological problems of language learners from collocations to other dimensions, such as semantic associations. Degree modifiers are adverbs that modify gradable words such as adjectives; adverbs of manner, place and time; and quantifiers. In previous studies it has been noted, first, that learners face problems with the frequency and distribution of degree modifiers. Granger (1998) and Hinkel (2002, 2003) have reported on clear overuse of certain items (e.g. *totally*, *completely*), others on underuse (e.g. *highly*; see Granger & Rayson 1998; also Granger 1998). Overuse is initiated because learners tend to favour certain modifiers at the expense of others: English language learners, for example, use *very*, *so*, *really* and *too* as so-called lexical teddy bears (see Pérez-Paredes & Díez-Bedmar 2012; Sabaté i Dalmaun & Curell i Gotorin 2007). Second, the phraseology of degree modifiers is also demanding of learners. Both Granger (1998) and Lorenz (1999) report on phraseological atypicalities, such as deviant collocations. These may be caused, for example, due to transfer from learners' mother tongue, but also because phraseological information in general is not available to learners in learning situations.

A recent analysis of degree modifiers in learner production (Jantunen 2015) also revealed that learners of Finnish face problems in adjective, adverb and quantifier modification. The study focused on the B1 level (i.e. the independent users' intermediate level) of language production described in the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001) within both foreign language and second language environments. Apart from the influence of learning context on learning and learner output, the findings support earlier observations on deviant phraseological combinations and distribution of degree modifiers in learner

writing. One clear finding was the overuse of the booster *oikein* (‘very’) in learner production in a foreign language learning context: it is favoured over the synonymous *hyvin* and *kovin*, which are typically more frequent than *oikein* in native written texts (Jantunen 2004; see also FLFNL).

A further contrastive analysis performed for the present article clarifies the picture of the usage of *oikein* in learner Finnish at the same CEFR level. Whereas the previous analysis of *oikein* (Jantunen 2015) concentrated solely on the frequencies (over- and underuse) of degree modifiers and only lightly touched on their phraseology, the present study tries to provide qualitative and causative information beyond these aspects. Concentrating on one single lexical item can provide information on the variety of difficulties learners face when acquiring lexicon, items belonging to a certain semantic field or syntactic set, or even when acquiring one single unit of meaning. The learner data come from the International Corpus of Learner Finnish (ICLFI, Jantunen & Pirkola 2015; Brunni et al. 2015) and the comparable native data from the Corpus of Translated Finnish (CTF, Mauranen 2000). The data include fictional as well as non-fictional texts, consisting of essays and argumentative text types as well as narratives and stories. Consequently, CTF broadly corresponds to ICLFI in the text types available. Texts vary, however, in aspects such as editing and producers: CTF data are edited and published texts produced by professional writers whereas ICLFI texts are not. This difference must be kept in mind when interpreting the results. The sizes of the sets of data are 342,000 tokens in ICLFI and 3,815,000 tokens in CTF. As in the previous analysis, the CEFR level is B1.

The analysis of the learner and native writing reveals, first of all, a general overuse of *oikein* among learners of Finnish. It is clearly more frequent in ICLFI than in the native data: in learners’ data the raw frequency is 416 and the normed value per 1 million tokens is 1216, whereas in native writing the frequencies are 290 and 76, respectively. That means that *oikein* is used 16 times more frequently by learners than it is by native language users. This result is in line with the general overuse of degree modifiers, especially boosters, in learner Finnish (see e.g. Jantunen 2015; for learner English, see Granger 1998; Lorenz 1999; Pérez-Paredes & Díez-Bedmar 2012; Sabaté i Dalmaun & Curell i Gotorin 2007).

Table 1 summarises the most frequent R1 collocates, i.e. the modified headwords, of *oikein* in both learner and native data. The data reveal that *oikein* is used as a modifier in cases that are non-frequent or non-existent in native data. Collocations, such as <OIKEIN PALJON> ‘very much’ and <OIKEIN KAUNIS> ‘very beautiful’, are examples of grammatically well-formed but phraseologically unusual co-occurrences and <degree modifier + adjective> and <degree modifier + quantifier> modification structures in learner data. Although they are comprehensible combinations, they violate the phraseological system of *oikein*, because it is rarely used as a modifier with KAUNIS and

PALJON by native speakers of Finnish (in written language). For example, the collocation <OIKEIN PALJON> does exist, in the CTF, although seldom. In four cases out of the total of six, it is used in compliments (e.g. *Kiitos oikein paljon* 'Thank you very much') in which it is anticipated. The association patterning for the collocation <OIKEIN PALJON> is therefore rather restricted in native data. Furthermore, the colligational pattern in which *oikein* would modify quantifiers in general are extremely rare in native data: the relative frequency per 1 million tokens is only 2.4 for the *oikein* <quantifier> colligation in native writing, whereas in learner writing it is as high as 160.8 (z = 4.73824, p = 0.00000) (see also Table 1 for *paljon* and *vähän*). Both sets of data share the frequent collocations <OIKEIN HYVIN> 'very well' and <OIKEIN HYVÄ> 'very good', which indicate that learners are familiar with some of the typical modification schemes of Finnish. Typical collocations such as <OIKEIN KUNNOLLA> and <OIKEIN KUNNON>, however, are unknown to the learners.

Table 1. The collocates (min. 5) of *oikein* in learner compared to native data, absolute and relative (per 1 million tokens) frequencies.

Learner data (*ICLFI*)	Native data (*CTF*)
PALJON 'much, a lot' 48 / 140.3	**HYVIN** 'well' 45 / 11.9
KAUNIS 'beautiful' 42 / 123.8	**HYVÄ** 'good' 38 / 10.1
HYVIN 'well' 33 / 96.5	KUNNOLLA 'hard, soundly' 14 / 3.7
HYVÄ 'good' 32 / 93.5	KUNNON 'good, hard' 10 / 2.7
MUKAVA 'nice' 21 / 61.4	**MUKAVA** 'nice' 10 / 2.7
HAUSKA 'funny' 13 / 3.0	KOVASTI 'hard, soundly', 6 / 1.6
KIVA 'nice' 11 / 32.2	MIELELLÄÄN 'gladly' 6 / 1.6
VAIKEA 'difficult' 10 / 29.2	TOSISSAAN 'seriously' 6 / 1.6
ISO 'big' 10 / 29.2	**PALJON** 'much, a lot' 6 / 1.6
PIENI 'small' 10 / 29.2	**HAUSKA** 'funny' 5 / 1.3
TÄRKEÄ 'important' 9 / 26.3	**KIVA** 'nice' 5 / 1.3
VIIHTYISÄ 'cosy' 8 / 23.4	
MIELENKIINTOINEN 'interesting' 7 / 20.5	
ONNELLINEN 'happy' 6 / 1.5	
ERILAINEN 'different' 5 / 14.6	
VÄHÄN 'little, few' 5 / 14.6	
VÄSYNYT 'tired' 5 / 14.6	

Nevertheless, the higher frequency of *oikein* in learner data does not necessarily indicate a higher collocational diversity (CD). In the present study, CD refers to the number of different collocates (here collocate lemmas) for the node *oikein*. Because the number of collocates is influenced by the frequency of the node (Dayrell 2007) and the corpus size (Stubbs 2001: 133; Jantunen 2004: 98–99), the collocates are counted using 150 concordance-line samples and represented in relation to the node frequency. That means that a higher ratio indicates a greater variation among collocates. In the learner data, the CD ratio is 37.6, which reflects that learners

tend to use rather fixed node−R1 collocate expressions; the CD ratio in native data is clearly higher, 50.3. There is clearly more variation in native writing than in learner writing regarding the collocations in the R1 position.

The modifier node *oikein* and its frequent R1 collocates form strings called bigrams, which are continuous and uninterrupted, grammatically complete or incomplete word sequences of two words (for *n*-grams, see e.g. Stubbs 2007; Scott & Tribble 2006; De Cock 2003; Salazar 2014). In other words, we can also claim that the variation in *oikein* bigrams is more fixed in learner writing. The question becomes if this also indicates a higher proportion of fixed expressions that are longer than bigrams. Table 2 lists the recurrent trigrams in both sets of data; only trigrams the frequency of which are at least 3 are taken into account to avoid writer idiosyncrasies. Table 2 clearly shows excessive use of lexical strings (i.e. less *n*-gram variation) in learner writing on level B1 in comparison to the native data: the share of left-hand trigrams (L2-L1-node) is 37.3% of all the concordance lines, for central trigrams (L1-node-R1) it is 50.5%, and for right-hand trigrams (node-R1-R2) it is 15.1%. In native data, trigrams do not play such a large role: the recurrent right-hand trigrams are non-existent, and the shares for others are under 10% (4.8% for left hand and 7.6% for central trigrams). Although some of the trigrams may well be generated due to tasks and writing guidelines (e.g. HUONE OLLA *oikein* [ROOM BE *oikein*], *oikein* KAUNIS LUONTO [*oikein* BEAUTIFUL NATURE]), the majority of the trigrams seems to suggest a less varied use of lexical patterns and repetition in learner writing. This is in line with the earlier findings concerning overuse of lexical bundles in non-native writing (see e.g. Salazar 2014; Paquot & Granger 2012: 138–139), which may reflect learners' more restricted lexical repertoire (Paquot & Granger 2012) and their tendency to use confident lexical and collocational teddy bears in writing (Granger 1998: 156) as well as L1 transfer (Paquot 2008, 2010).

Table 2. Trigrams (min. 3) in learner and native data and their frequency

Learner data (*ICLFI*)	Native data (*CTF*)
Left-hand trigrams:	Left-hand trigrams:
SE OLLA *oikein* (IT BE *oikein*) 33	EI OLLA *oikein* (NO BE *oikein*) 4
ME OLLA *oikein* (WE BE *oikein*) 16	SE OLLA *oikein* (IT BE *oikein*) 4
MINÄ OLLA *oikein* (I BE *oikein*) 16	TÄMÄ OLLA *oikein* (THIS BE *oikein*) 3
HÄN OLLA *oikein* (HE BE *oikein*) 13	KUN OLLA *oikein* (WHEN BE *oikein*) 3 (14)
HUONE OLLA *oikein* (ROOM BE *oikein*) 9	
SIELLÄ OLLA *oikein* (THERE BE *oikein*) 8	Central trigrams:
JA OLLA *oikein* (AND BE *oikein*) 7	OLLA *oikein* HYVÄ (BE *oikein* GOOD) 8
SUOMI OLLA *oikein* (FINLAND BE *oikein*) 4	OLLA *oikein* MUKAVA (BE *oikein* NICE) 6
ISÄ OLLA *oikein* (DAD BE *oikein*) 4	OLLA *oikein* HAUSKA (BE *oikein* FUNNY) 5
OLLA MYÖS *oikein* (BE ALSO *oikein*) 3	OLLA *oikein* KIVA (BE *oikein* NICE) 3 (22)
MUTTA OLLA *oikein* (BUT BE *oikein*) 3	
OLLA MINÄ *oikein* (BE I *oikein*) 3	Right-hand trigrams:

PERHE OLLA *oikein* (FAMILY BE *oikein*) 3 TÄÄLLÄ OLLA *oikein* (HERE BE *oikein*) 3 (155) Central trigrams: OLLA *oikein* KAUNIS (BE *oikein* BEAUTIFUL) 33 OLLA *oikein* HYVÄ (BE *oikein* GOOD) 21 OLLA *oikein* MUKAVA (BE *oikein* NICE) 18 OLLA *oikein* HAUSKA (BE *oikein* FUNNY) 11 PITÄÄ *oikein* PALJON (LIKE *oikein* MUCH) 11 OLLA *oikein* KIVA (BE *oikein* NICE) 10 OLLA *oikein* VAIKEA (BE *oikein* DIFFICULT) 10 OLLA *oikein* PALJON (BE *oikein* MUCH) 9 OLLA *oikein* PIENI (BE *oikein* SMALL) 9 OLLA *oikein* ISO (BE *oikein* BIG) 9 OLLA *oikein* TÄRKEÄ (BE *oikein* IMPORTANT) 7 OLLA *oikein* HYVIN (BE *oikein* WELL) 6 OLLA *oikein* MIELENKIINTOINEN (BE *oikein* INTERESTING) 6 OLLA *oikein* VIIHTYISÄ (BE *oikein* COSY) 6 OLLA *oikein* ONNELLINEN (BE *oikein* HAPPY) 6 PUHUA *oikein* HYVIN (SPEAK *oikein* WELL) 5 OLLA *oikein* VÄSYNYT (BE *oikein* TIRED) 5 JA *oikein* KAUNIS (AND *oikein* BEAUTIFUL) 4 OLLA *oikein* KYLMÄ (BE *oikein* COLD) 4 OLLA *oikein* KIINNOSTAVA (BE *oikein* INTERESTING) 4 OLLA *oikein* ERILAINEN (BE *oikein* DIFFERENT) 4 OLLA *oikein* PITKÄ (BE *oikein* TALL, LONG) 3 OSATA *oikein* HYVIN (KNOW, CAN *oikein* WELL) 3 VIIHTYÄ *oikein* HYVIN (FEEL COMFORTABLE *oikein* WELL) 3 SE *oikein* PALJON (IT *oikein* MUCH) 3 (210) Right-hand trigrams: *oikein* MUKAVA JA (*oikein* NICE AND) 12 *oikein* KAUNIS JA (*oikein* BEAUTIFUL AND) 6 *oikein* PALJON JA (*oikein* MUCH AND) 6 *oikein* KAUNIS LUONTO (*oikein* BEAUTIFUL NATURE) 6 *oikein* KAUNIS PAIKKA (*oikein* BEAUTIFUL PLACE) 4 *oikein* HYVIN ENGLANTI (*oikein* WELL ENGLISH) 4 *oikein* PALJON TYÖ (*oikein* MUCH WORK) 4 *oikein* PALJON AIKA (*oikein* MUCH TIME) 3 *oikein* MIELENKIINTOINEN JA (*oikein* INTERESTING AND) 3 *oikein* KIVA JA (*oikein* NICE AND) 3 *oikein* HYVÄ JA (*oikein* GOOD AND) 3 *oikein* ISO JA (*oikein* BIG AND) 3 *oikein* HYVIN VIRO (*oikein* WELL ESTONIAN) 3 *oikein* KAUNIS KAUPUNKI (*oikein* BEAUTIFUL TOWN) 3 (63)	-

In the previous study on the phraseology of synonymous boosters (Jantunen 2004), it became clear that *oikein* indicates a positive semantic prosody. Semantic prosody refers to a discourse or

pragmatic function of an item, or a certain kind of aura or attitude (negative or positive) associated with the word (core) in question (see e.g. Sinclair 1996; Louw 2000; Hunston 2007). In Jantunen's (2004) study, 70% of the occurrences of *oikein* provided evidence of positive evaluation, and only 28% showed negative associations. In the present set of native data (which is larger than in Jantunen 2004), the shares are similar: in the analysis of each of the 416 concordance lines it turned out that 68% of the occurrences of *oikein* found in the cotext show positive prosody, and only 28% were used in a negatively loaded context. In learner data, on the other hand, the prosodies differ, with the share of positive prosody being slightly larger (76%; $z = 2.54944$, $p = 0.01078$) and the share of negative prosody smaller (21%; $z = 2.43475$, $p = 0.0151$). It seems that learners of Finnish emphasise the positive quality of *oikein* in their writing. Although the share of negative prosody is slightly smaller in the learner data, the actual co-occurrences that indicate negative evaluation stick out. The data reveal that learners tend to use *oikein* with words whose meaning is clearly negative and, moreover, which are usually modified with other boosters in native Finnish. These words are, for example, VAIKEA 'difficult' (see also Table 1 above), VÄSYNYT 'tired', SURULLINEN 'sad', VIHAINEN 'angry' and SAIRAS 'ill'. In native writing, these typically are modified with booster such as *hyvin*, *erittäin*, *kovin* and *tosi*, and only occasionally with *oikein*. Another difference between learner and native writing is that learners have a preference for using *oikein* in cases in which the head word refers to a negatively evaluative characteristic (e.g. ANKARA 'strict, ILKEÄ 'mean', LAISKA 'lazy'), emotional state (e.g. SURULLINEN 'sad', ONNETON 'unhappy', VIHAINEN 'angry') or physical state (e.g. VÄSYNYT 'tired', SAIRAS 'ill', NÄLKÄINEN 'hungry') of a person.

A detailed analysis of the semantic preferences of *oikein* in learner data shows that it often occurs in cotexts where a person is characterised or described. There are numerous examples where it is used with various collocates that denote the qualities of a person. The share of the semantic preference for 'personality' in learner data is 12.5%, whereas in native data the share is only 5.5% ($z = 3.0939$, $p = 0.002$). This preference manifests itself, for example, as collocates YSTÄVÄLLINEN 'friendly', YLPEÄ 'proud', UTELIAS 'nosy' and ANKARA 'strict'), see examples (1)–(4).

(1) *Hän on oikea [pro oikein]* ***ystävallinen*** *[pro ystävällinen], pitää urheilusta ja juhliasta [pro juhlista]--.*
'S/he is very friendly, likes sports and parties --.'

(2) *Minä olen oikein* ***ylpeä*** *jälkeenpäin.*
'I'm very proud afterwards.'

(3) *Se on* ***ystävällinen*** *ja oikein* ***utelias****.*
'It is friendly and very nosy.'

(4) *Opettaja on oikein **ankara** ja konservatiivinen.*
'The teacher is very strict and conservative.'

The collocates to the right also show that *oikein* indicates another difference in its semantic associations: the share of semantic preference for 'feeling, emotion' is 6.7% in learner data, but in native data it is only 1.7% ($z = 3.1005$, $p = 0.00194$). In these cases, *oikein* is associated with collocates such as TYYTYVÄINEN 'satisfied', SURULLINEN 'sad' and ONNELLINEN 'happy' in the learner data (5)–(7).

(5) *Minä olin tietysti oikein **tyytyväinen** siihen tilanteeseen.*
'I was of course very satisfied with that situation.'

(6) *Olin oikein **surullinen** ja luulin että hän oli yksikertailtä [pro yksinkertaisesti] tyhmä.*
'I was very sad and I thought that s/he was simply stupid.'

(7) *Olin sillä neljä viikkoa – leväsin [pro lepäsin], matkailin ja olin oikein **onnellinen**.*
'I was there for four weeks – I rested, travelled and was very happy.'

The most notable difference between these two sets of data is, however, clearly in the occurrence of semantic preference for 'intensity' in the cotext. Contrary to the previous cases, this preference appears significantly often in native data (20.0% of the occurrences of *oikein*), but in learner data it is extremely rare (1.0%) ($z = 8.7931$, $p = 0.00000$). In the native data, in the immediate environment of *oikein*, for example, the following collocates are often found (note that these are non-existent in learner data, see also Table 1): KUNNON 'good, hard', KUNNOLLA 'badly, soundly', KOVASTI 'hard, soundly', TOSISSAAN 'serious, seriously'; see examples (8)–(11). It seems that learners of Finnish deviate from the phraseological system of *oikein* by overusing certain preferences and underusing, in particular, the preference for 'intensity'.

(8) *Jossain vaiheessa aika myöhään tein oikein **kunnon** lenkin.*
'At some point rather late I had a very good jog.'

(9) *Jymäytin sitä [ovi] pari kertaa oikein **kunnolla**.*
'I beat it [door] twice very soundly.'

(10) *Menestyvä liikemieskin voi olla rehellinen, jos oikein **kovasti** yrittää, --*
'A successful businessman can be honest, if s/he tries very hard, --'

(11) *Hän hytisi, häntä paleli nyt oikein **tosissaan**, ---*
' S/he shivered, s/he was freezing very seriously now, --'

4. Dictionaries, phraseology and *oikein*

Although the roots of computer-based dictionaries date back to the 1960s and the first dictionaries based on lexical databases were launched in the late 1970s and 1980s (Granger 2012: 1), the integration of corpus data, phraseology analysis and dictionary design have taken place relatively recently. And even though Hanks (2012: 63) states that "[t]he impact of corpus data on synchronic lexicography since 1987 (--) has been overwhelming", we must keep in mind that this concerns only a small number of the world's languages. Furthermore, the changes in corpora have made things different. According to Granger (2012: 3), the analysis of raw corpus data is no longer the primary manner to build dictionary entries: large, billion-word corpora are usually analysed (annotated) before they are used by lexicographers; this helps the dictionary writers to more easily find more exact and better real-language-oriented patterns for entries. Phraseological relations, however, are still seldom annotated in the data, though they can be found using corpus tools such as concordancers. But this process can be time-consuming for each word entry. This final section discusses the presentation and role of phraseology in dictionaries.

In dictionary making, corpus evidence can be used not only as a source for examples but, despite its time-consuming nature, to locate the conventional phraseology of a certain register, text type or language. Hanks (2012: 64) even states that, in the future, phraseological norms and the contextualisation of words will become increasingly important in the design of dictionary entries. Here the situation of Finnish language dictionaries should be considered. They do not provide explicit phraseological evidence. Instead, they describe the meaning and usage of words, giving short descriptions, synonyms and (real language) examples. The phraseological information can only be found if the user is able to analyse the given examples and if the examples provide information on real language usage. The following example comes from the *Dictionary of Contemporary Finnish (DCF)*, which is a dictionary of standard Finnish that aims to describe Finnish in general. It is more a monolingual dictionary for native speakers to find information on word meanings than it is a dictionary for language learners on how to use words. It should be mentioned, however, that entries are not based solely on intuition, but on a growing word archive of contemporary Finnish (see the *DCF*). The entry for *oikein* in the *DCF* defines the degree modifier as follows:

oikein
erittäin, hyvin, kovin, sangen, perin, varsin. *Oikein hyvä! Katseli oikein tarkkaan. Nukuin oikein hyvin. Oikein paljon kiitoksia.*
('Very good! Looked at very carefully. I slept very well. Thank you very much.')

First, the entry lists several synonymous degree modifiers for *oikein* (all meaning roughly 'very'). *Oikein* is defined using its synonyms, and no proper definition is provided. The listing of

synonyms may also suggest that any of these synonymous modifiers are interchangeable in any context, because any register or style differences for these synonymous items are not given in the dictionary entry. Of course, this is not the case (see e.g. Jantunen 2004). Second, the example sentences (which are typically from real contexts) give some phraseological information of the degree modifier: the reader might assume that *oikein* collocates, for example, with HYVÄ 'good', HYVIN 'well', TARKKAAN 'carefully' and that it is used in compliments, such as *oikein paljon kiitoksia* 'thank you very much'. This phraseological information is valuable for dictionary user but is not provided in an explicit manner. The dictionary user must draw conclusions regarding the phraseology from a few example sentences. Information on the phraseological usage of *oikein* is implicit and not clearly explained for dictionary users. As stated above, if the phraseological patterns were provided, it would require an enormous amount of work due to the huge amount of data that must be analysed before writing an entry. However, as Hanks (2012: 82) suggests, "all serious future lexicography will be corpus-driven, no longer based merely on collections of citations and certainly not merely a matter of guesswork based on the speculation". The conclusion here is that corpus analyses, perhaps more and more automated, will be increasingly needed in Finnish language dictionary making.

Online learner dictionaries are one application that benefits from the usage of corpus data and research, but corpus-based learner dictionaries as well as other corpus-based pedagogical materials have been rare in Finnish. Hanks (2012: 62), however, points out that in English, "the first major impact of corpora on lexicography was on a dictionary for foreign learners, namely *COBUILD*", which actually dates back to the end of the 1980s. For Finnish, such dictionaries did not exist before the pilot project *ConLexis* (Jantunen et al. 2013), which is an online corpus-based learner dictionary of synonymous items. It has been produced in a pilot project,[1] the aim of which has been to combine phraseological information with online dictionary entries. Although both Rundell (1998) and Järventausta (2009) state that learner dictionaries contain wider descriptions of phraseology than general dictionaries, this is not the case in Finnish learner dictionaries (especially when phraseology is understood from a Sinclairian perspective). The principles behind the *ConLexis* dictionary are based on the thoughts of Sinclair and Renouf (1988), Atkins (1993) and Järventausta (2009) regarding learner dictionary compilation. In their views, learner dictionaries should do the following:

- Give detailed information on style, genre and cotextual usage (phraseology).
- Explain high-frequency core vocabulary.

[1] The *ConLexis* pilot project has been conducted in cooperation with the University of Oulu and the Oulu Adult Education Centre and funded by the Finnish National Board of Education.

- Explain associations with other expressions (synonyms, antonyms, hyponyms).
- Explain the core inflectional forms.
- Have simple and plain definitions.
- Be based on a large amount of electronic data, which could be used in writing the definitions.
- Give frequencies of items, patterns, meanings and collocations.
- Give qualitative information of grammatical and phraseological patterns.
- Provide real-world language examples.
- Explain the (phraseological) differences of synonyms and antonyms.
- Provide information on style, registers and text types.

ConLexis is a monolingual, free-of-charge dictionary targeting Finnish language learners on the B1 CEFR level or higher. At the moment it contains about 300 entries. The starting point in lexeme selection was the frequency of words: a frequency list of contemporary Finnish was used to reveal the core items in Finnish (and also to provide the frequency information for the dictionary). However, rare words have also been included, because the dictionary focuses on synonyms and antonyms, which has led to the description of several low-frequency but synonymous or antonymous items. Information represented in the *ConLexis* entries is based on time-consuming corpus analyses and has a clear phraseological emphasis. Each entry in the dictionary explains the word alongside its definition and inflection as well as provides links to synonyms and antonyms and a frequency-based list of the most important collocates. For each collocate, a list of example phrases derived from the corpora is also presented. Furthermore, colligational associations are provided together with example concordances.

The entries have been written based on the Finnish Textbank, which consists of texts from the 1990s. Occasionally, the Internet has also been used to provide examples of up-to-date information on language usage. The collocational associations of the items have been analysed using the web-based tool www-Lemmie2,[2] which allows the user to obtain word frequencies, collocations and concordances. The word entries include definitions, inflection tables, collocations, *n*-grams, synonyms, antonyms, and in some cases further information. The dictionary provides paradigmatic information (synonyms, antonyms, inflection) and syntagmatic information (phraseology and cotexts). Examples are authentic but simplified when necessary, and they are displayed visually as

[2] http://metashare.csc.fi/repository/browse/www-lemmie/aff491b8fccc11e18b49005056be118e2f69c385f23b4ad0a8042a073d009f4d/

concordances. *ConLexis* is not normative. Instead, it tells how words are typically used in natural target language, providing clear information on phraseological conventions. An example of the entry for *oikein* is displayed below.

oikein (adv.)
Oikein vahvistaa tai tehostaa seuraavan saman merkitystä. Se kertoo, että jotakin on enemmän kuin tavallisesti. Yleensä se vahvistaa adjektiivia tai adverbia. Sen käyttö on muita määritteitä rajoittuneempaa, sillä sitä käytetään erityisesti adjektiivien *hyvä*, *mukava* ja *tyytyväinen* sekä adverbien *hyvin*, *kunnolla* ja *tosissaan* määritteenä. Se voi määrittää muitakin sanoja, mutta se ei ole kovin tavallista.

['*Oikein* strengthens or enhances the meaning of the following word. It indicates that something is more than usual. It generally reinforces the meaning of an adjective or adverb. Its use is more limited than other degree modifiers, because it is used particularly as a modifier for the adjectives *good*, *nice* and *satisfied* and the adverbs *well*, *properly* and *seriously*. It can modify other words, but this practice is not very common.']

Vierussanat [collocates]

- *hyvin* (Each collocate is linked to a collection of examples, below is an example of the collocate *hyvin.*)
- *hyvä*
- *kunnolla*
- *mukava*
- *tosissaan*
- *tyytyväinen*

Viihdymme täällä	**oikein hyvin.**	
Kyllä ruotsia piti melko tavalla harjoitella, mutta hommat sujuvat nyt jo	**oikein hyvin.**	
Mitään erikoista sairautta ei ole ollut, joten hän on jaksanut	**oikein hyvin.**	
Kansainvälinen loppu poliittiselle uralle sopii tietenkin	**oikein hyvin,**	kun olen koko ajan olen ollut niin kansainvälinen.
Suomessa tiedetään	**oikein hyvin,**	miten kalliita huippumodernit lentokoneet ovat.
Ymmärrän	**oikein hyvin,**	että seksuaalikasvatus on muutakin kuin lapsenteko-oppia.
Tänään onnistuin kahdesti	**oikein hyvin,**	sympaattinen Schmitt mietti.
Bennille sopii	**oikein hyvin,**	että teokset liitetään tieteiskirjallisuuteen.
Viime yönä nukuin	**oikein hyvin,**	kahdeksan tuntia.
Hän on oppinut esimerkiksi englantia	**oikein hyvin,**	mutta matematiikan sanallisissa tehtävissä hänellä on vaikeuksia.

The entry continues with a list and links to synonymous degree modifiers as well as information on antonyms:

Synonyymit [synonyms]

erittäin, hyvin, varsin, tosi, todella, sangen; *myös kovin, järin*

Antonyymit [antonyms]

Tällä sanalla ei ole selviä vastakohtia. Adjektiivin, adverbin ja kvanttorin vähäistä määrää ilmaisevat *aika*, *melko*, *jokseenkin*, *kohtalaisen*, *suhteellisen*, *verrattain*.

['This word has no clear antonyms. The low degree of the meaning of adjective, adverb and quantifiers are expressed using fairly, rather, somewhat']

Lisätietoa [further information]

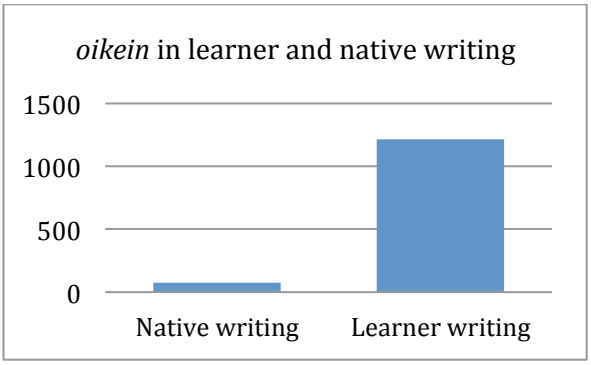
Astemääritteet, kuten *hyvin* ja *kovin*, voivat määrittää hyvin monenlaisia sanoja ja ne määrittävät usein keskenään myös samoja sanoja. Tärkein erottava tekijä on, että *kovin* ja *järin* ovat kieltolauseessa (eli lauseessa on kieltosana *ei*). Muita käytetään myönteisessä lauseessa.

['Degree modifiers such as *hyvin* and *kovin* can modify a wide range of words, and they often modify the same words. The main distinguishing feature is that *kovin* and *järin* are used in negative sentences (i.e. there is a negative verb in the sentence). Others are used in positive sentences.']

ConLexis is based on native language data only, but the *Macmillan English Dictionary* (second edition, 2007) also contains data from learner language corpora, which gives the dictionary a more extensive dimension.[3] The 50-page appendix called 'Improve your writing skills' provides corpus-based evidence with a particular focus on academic or professional writing from non-native and native perspectives. This section is the outcome of a collaborative project between Macmillan and the Centre for English Corpus Linguistics (Université catholique de Louvain), in which the International Corpus of Learner English (ICLE) is used as source for learner production (for more on the project, see De Cock & Paquot 2010.) It includes, for example, information boxes called 'Get it right', 'Be careful', and 'Collocations', which all provide examples of how learners tend to use certain items and how those items are used in native written language. Using these Macmillan notes on language usage as a model, the entry for *oikein* could have the following information boxes (Figure 1) on its frequency and phraseological usage in *ConLexis* or another learner dictionary. These boxes could be provided for learners of Finnish, especially in online dictionaries. (These observations are based on the above analysis of *oikein*, the Finnish Textbank and *ConLexis*, and they should not be taken to concern written Finnish as a whole.)

Learners of Finnish often use the degree modifier *oikein*. Although it is used in written language, it is collocationally restricted, and in many cases *hyvin* and *erittäin* can be used instead.

[3] Longman's dictionaries for language learners are also based on learner data, namely the publisher's own Longman Learner Corpus (see e.g. Nesselhauf 2004). A more comprehensive reference tool for learners of English is the *Louvain English for Academic Purposes Dictionary* (see Granger & Paquot 2015).



Collocation
Adjectives frequently used with *oikein*:
hyvä, mukava, tyytyväinen
Se on oikein hyvä lounaspaikka.

Adverbs frequently used with *oikein*:
hyvin, kunnolla, tosissaan
Hän suuttui oikein kunnolla.

Figure 1. 'Be careful' and collocation information boxes, based on the *Macmillan English Dictionary*.

5. Conclusions

Language teaching is currently in a state of change. Technology and teaching meet in several ways and situations, and together they offer interesting pathways to better learning. Although it has been stated that technology and corpus data are still too seldom used in teaching, learning, and material production in many languages, the situation is changing rapidly. The existence and capability to use electronic devices alone creates a demand for new methods. Because a vast number of corpora from several languages are available and more user-friendly corpus tools have come to the market, there is no longer any excuse to disregard teaching and learning methods based on natural language or conventional phraseology of the target language in teaching materials and (learner) dictionaries. For learners, acquiring new words in isolation of the usage contexts and trying to guess the correct and context-appropriate ways of using them is a task that is both old-fashioned and frustrating. Instead, we ought to strengthen learners' existing and developing phraseology (see Vetchinnikova 2014). Such a rapid transformation, however, would be challenging. There would be much to change, starting from the education of language teachers and dictionary writers. In English language teacher education, this area has already seen significant progress. However, despite its record of producing skilled users of Finnish, the education of Finnish language teachers still possesses a lack of data-driven and phraseology-based teaching and learning methods, including in learning materials such as learner dictionaries. Furthermore, Finnish learner dictionaries seem to meet users' receptive needs only; the use of information that would help learners with their productive needs remains in its infancy in dictionaries and other writing aids. Therefore, there is also much that needs to be done. One starting point lies in the systematic improvement and supplementing of learner language and native corpora, their analysis, the integration of natural language and phraseology into language teacher education, and data-driven lexicography.

References

Ahmadian, Moussa, Hooshang Yazdani & Ali Darabi 2011. Assessing English Learners' Knowledge of Semantic Prosody through a Corpus-Driven Design of Semantic Prosody Test. English Language Teaching 4(4): 288–298. http://www.ccsenet.org/journal/index.php/elt/article/view/13385

Atkins, B. T. Sue 1993. Theoretical lexicography and its relation to dictionary-making. *Dictionaries: Journal of the Dictionary Society of North* 14: 4–43.

Brunni, Sisko, Liisa-Maria Lehto, Jarmo H. Jantunen & Valtteri Airaksinen 2015. How to annotate morphologically rich learner language. Principles, problems and solutions. *Learner Corpus Research: LCR2013 Conference Proceedings*. Bergen Language and Linguistics Studies 6, eds. Ann-Kristin Helland Gujord, Susan Nacey & Silje Ragnhildstveit, 133–152. Bergen: University of Bergen. https://bells.uib.no/bells/article/view/812

CEFR = Council of Europe 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Chambers, Angela 2007. Popularising corpus consultation by language learners and teachers. *Corpora in the foreign language classroom*, eds. Encarnación Hidalgo, Luis Quereda & Juan Santana, 3–16. Amsterdam: Rodopi.

ConLexis. Online corpus-based dictionary for learners of Finnish*,* eds. Jantunen, Jarmo H., Tanja Tammimies, Marjo Kumpulainen & Teemu Tokola. http://wiki.virtues.fi/conlexis/

Dayrell, Carmen 2007. A quantitative approach to investigate collocations in translated texts'. *International Journal of Corpus Linguistics*, 12(3): 415–444.

DCF = *The Dictionary of Contemporary Finnish*. *Meta-Share*. http://metashare.csc.fi/repository/browse/dictionary-of-contemporary-finnish/a6288b88754e11e4a157005056be118e278c5746fc1f420181f255f6abeb6d86/)

De Cock, Sylvie 2003. *Recurrent Sequences of Words in Native Speaker and Advanced Learner Spoken and Written English*. PhD dissertation. Louvain-la-Neuve: Centre English Corpus Linguistics, Catholique University Louvain.

De Cock, Sylvie & Magali Paquot 2010. The monolingual learners' dictionary as a productive tool: the contribution of learner corpora. *Corpus-based approaches to English Language Teaching*, eds. Mari Carmen Campoy, Begoña Bellés-Fortuno & Ma Lluisa Gea-Valor, 195–204. London: Continuum.

FLFNL = Frequency Lexicon of the Finnish Newspaper Language 2004. *The Language Bank of Finland.* http://metashare.csc.fi/repository/browse/frequency-lexicon-of-the-finnish-newspaper-language/28c4b13a2cc111e2a376005056be118e2923899d05fe4eac9b1ed7762e344166/

Granger, Sylviane 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. *Phraseology: Theory, Analysis and Applications*, ed. Anthony P. Cowie, 145–160. Louvain-la-Neuve: Centre English Corpus Linguistics, Catholic University Louvain.

Granger, Sylviane 2012. Introduction: Electronic lexicography – from challenge to opportunity. *Electronic Lexicography*, eds. Sylviane Granger & Magali Paquot, 1–12. Oxford: Oxford University Press.

Granger, Sylviane 2015. Contrastive interlanguage analysis. A reappraisal. *International Journal of Learner Corpus Research* 1(1): 7–24.

Granger, Sylviane & Fanny Meunier 2008. Phraseology in language learning and teaching. Where to from here? *Phraseology in Foreign Language Learning and Teaching*, eds. Fanny Meunier & Sylviane Granger, 247–252. Amsterdam: John Benjamins.

Granger, Sylviane & Magali Paquot 2015. Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid. *Lexicographica* 31(1): 118–141.

Granger, Sylviane & Paul Rayson 1998. Automatic profiling of learner texts. *Learner English on Computer*, ed. Sylviane Granger, 119–131. London: Longman,

Hanks, Patrick 2012. Corpus evidence and electronic lexicography. *Electronic Lexicography*, eds. Sylviane Granger & Magali Paquot, 57–82. Oxford: Oxford University Press.

Hinkel, Eli 2002. Second Language Writers' Text. Linguistic and Rhetorical Features. Mahwah NJ: Erlbaum.

Hinkel, Eli 2003. Adverbial markers and tone in L1 and L2 students' writing. *Journal of Pragmatics* 35(7), 1049–1068. http://dx.doi.org/10.1016/S0378-2166(02)00133-9

Hoey, Michael 1997. Hoey, M. 1997. From concordance to text structure: new uses for computer corpora. – Barbara Lewandowska-Tomaszczyk & Patrick Melia (eds.) *PALC'97. Applications in Language Corpora Proceedings*, 2–23. Lódz University Press.

Hunston, Susan. 2007. Semantic prosody revisited. *International Journal of Corpus Linguistics* 12(2), 249–268.

Jantunen, Jarmo H. 2004. *Synonymia ja käännössuomi: korpusnäkökulma samamerkityksisyyden kontekstuaalisuuteen ja käännöskielen leksikaalisiin erityispiirteisiin*. Joensuun yliopiston humanistisia julkaisuja 35. [Synonymity and Translated Finnish. A Corpus-based View of Contextuality of Synonymous Expressions and Lexical Features Specific to Translated language] http://epublications.uef.fi/pub/urn_isbn_952-458-479-4/urn_isbn_952-458-479-4.pdf

Jantunen, Jarmo H. 2009. "Minulla on aivan paljon rahaa." Fraseologiset yksiköt suomen kielen opetuksessa. ["I have really lots of money." Phraseological units in the teaching of Finnish]. *Virittäjä* 113(3): 356–381.

Jantunen, Jarmo H. 2015. Oppimiskontekstin vaikutus oppijanpragmatiikkaan: astemääritteet leksikaalisina nallekarhuina. [Learning context and its effect on learner pragmatics: degree modifiers as lexical teddy bears.] *Lähivõrdlusi. Lähivertailuja* 25: 105–136.

Jantunen, Jarmo Harri & Sisko Brunni 2013. Morphology, lexical priming and second language acquisition: a corpus-study on learner Finnish. – Sylviane Granger, Gaetanelle Gilquin & Fanny Meunier (eds.) *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead*. Corpora and Language in Use 1, 235–246. Louvain-la-Neuve: Presses Universitaires de Louvain.

Jantunen, Jarmo H., Marjo Kumpulainen, Tanja Tammimies & Teemu Tokola 2013. Korpuspohjaista oppijansanakirjaa tekemässä: esimerkkinä ConLexis [Towards a corpus-based online learner dictionary: ConLexis]. *Lähivõrdlusi. Lähivertailuja* 23: 89–120.

Jantunen, Jarmo Harri & Silja Pirkola 2015. Oppijansuomen sähköiset tutkimusaineistot: nykytilanne. *Virittäjä* 119(1): 88–103.

Järventausta, Marja 2009. Kakkossuomen perussanakirja. *Virittäjä* 113(1): 89–100.

Kjellmer, Göran 1991. A Mint of Phrases. *Corpus Linguistics: Studies in Honour of Jan Svartvik*, eds. Karin Aijmer & Bengt Altenberg, 111–127. London: Longman.

Lorenz, Gunter R. 1999. *Adjective Intensification: Learners Versus Native Speakers. A Corpus Study of Argumentative Writing*. Language and Computers: Studies in Practical Linguistics 27. Amsterdam: Rodopi.

MacMillan English Dictionary for Advanced Learners 2007. Second edition. Oxford: MacMillan Education.

Louw, Bill. 2000. Contextual prosodic theory: Bringing semantic prosodies to life. *Words in context. In honour of John Sinclair*, eds. Chris Heffer & Helen Sauntson, 48–94. Birmingham: ELR.

Mauranen, Anna 2000. Strange strings in translated language: A study on corpora. *Intercultural Faultlines: Research Models in Translation Studies I. Textual and Cognitive Aspects*, ed. Maeve Olohan, 119–141. Manchester: St. Jerome.

Meunier, Fanny 2011. Corpus linguistics and second/foreign language learning: exploring multiple paths. *RBLA* 11(2): 459–477.

Nesselhauf, Nadja 2004. Learner corpora and their potential for language teaching. *How to use corpora in language teaching*, ed. John M. Sinclair, 125–152. Amsterdam: Benjamins.

Nesselhauf, Nadja 2005. *Collocations in a Learner Corpus*. Studies in Corpus Linguistics 14. Amsterdam: John Benjamins.

Nesselhauf, Nadja 2009. Co-selection phenomena across new Englishes: Parallels (and differences) to foreign learner varieties. *English World-Wide* 30,1–26.

Paquot, Magali 2008. Exemplification in learner writing: A cross-linguistic perspective. *Phraseology in Foreign Language Learning and Teaching*, eds. Fanny Meunier & Sylviane Granger, 101–119. Amsterdam: John Benjamins.

Paquot, Magali 2010. *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London: Continuum.

Paquot, Magali & Sylviane Granger 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics* 32: 130–149.

Pérez-Paredes, Pascual, María Belén Díez-Bedmar 2012. The use of intensifying adverbs in learner writing. *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*, eds. Yukio Tono, Yuji Kawaguchi & Makoto Minegishi, 105–124. Amsterdam: John Benjamins

Rundell, Michael 1998. Recent trends in English pedagogical lexicography. *International Journal of Lexicography* 11(4): 315–342.

Römer, Ute 2008. Corpora and language teaching. *Corpus Linguistics. An International Handbook* (volume 1), eds. Anke Lüdeling & Merja Kytö, 112–130. Berlin: Mouton de Gruyter.

Sabaté i Dalmau, Maria, Hortènsia Curell i Gotor 2007. From 'Sorry very much' to 'I'm ever so sorry'. Acquisitional patterns in L2 apologies by Catalan learners of English. *Intercultural Pragmatics* 4(2), 287–315.

Salazar, Danica 2014. *Lexical Bundles in Native and Non-native Scientific Writing. Applying a corpus-based study to language teaching*. Studies in Corpus Linguistics 65. Amsterdam: John Benjamins

Scott, Mike & Christopher Tribble 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Studies in Corpus Linguistics 22. Amsterdam: John Benjamins.

Sinclair, John M. 1987. Collocation: a progress report. *Language Topics. Essays in Honour of Michael Halliday*, eds. Ross Steele & Terry Threadgold, 319–331. Amsterdam: Benjamins.

Sinclair, John M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John M. 1996. The search for units of meaning. *Textus* IX: 75–106.

Sinclair, John M. & Antoinette Renouf. 1988. A lexical syllabus for language teaching. *Vocabulary and Language Teaching*, eds. Ronald Carter & Michael McCarthy, 140–158. London: Longman.

Stewart, Dominic 2010. *Semantic Prosody: A Critical Evaluation*. New York & London: Routledge.

Stubbs, Michael 1995. Collocations and semantic profiles: On the *cause* of trouble with quantitative studies. *Functions of Language* 2(1): 23–55.

Stubbs, Michael 2007. Quantitative data on multiword sequences in English: The case of the word 'world'. *Text, Discourse and Corpora: Theory and Analysis*, eds. Michael Hoey, Michaela Mahlberg, Michael Stubbs & Wolfgang Teubert, 163–189. London: Continuum.

Thewissen, Jennifer 2008. The phraseological errors of French-, German-, and Spanish speaking EFL learners: Evidence from an error-tagged learner corpus. *Proceedings from the 8th Teaching and Language Corpora Conference* (*TaLC 8*), eds. Ana Frankenberg-Garcia, Tawfiq Rkibi, Maria do Rosário Braga da Cruz, Ricardo Carvalho, Cristina Direito & Diogo Santos-Rosa, 300–306. Lisboa: Associação de Estudos e de Investigação Científica do ISLA-Lisboa.

Vetchinnikova, Svetlana 2014. *Second language lexis and the idiom principle*. Helsinki: University of Helsinki. https://helda.helsinki.fi/handle/10138/135691

Wang, H. & Wang, T. 2005. A contrastive study on the semantic prosody of CAUSE. *Modern Foreign Language.* 28(3): 297–307.

Wei, Naixing 2006. A Corpus-based Contrastive Study of Semantic Prosodies in Learner English. *Foreign Language Research*, 132: 50–54.

Wray, Alison 2002. *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

Wray, Alison & Michael R. Perkins 2000. The functions of formulaic language: an integrated model. *Language and Communication* 20: 1–28.

Xiao, Richard & Tony McEnery 2006. Collocation, Semantic Prosody, and Near Synonymy: A Cross-Linguistic Perspective. *Applied Linguistics* 27(1): 103–129.