A new approach to measuring lexical sophistication in L2 oral production

Christina Lindqvist*, Anna Gudmundson** and Camilla Bardel**
*Uppsala University, **Stockholm University

The aims of this chapter are a) to give a comprehensive description of a new tool for lexical profiling by reporting how it was developed, and b) to indicate possible areas of use and future developments of the tool. The tool has been used for measuring the lexical sophistication of Swedish learners of French and Italian. The different steps of development have partly been presented in previous studies (Bardel & Lindqvist, 2011; Bardel, Gudmundson & Lindqvist, 2012; Lindqvist, Bardel & Gudmundson, 2011) but are complemented here through a detailed account of the tool, in order to enable replication and use of the method with other languages. The outline of this chapter is as follows: first, as a background, we provide a survey of methods designed to measure lexical richness in L2 production. Then we discuss the inherent differences between written and spoken language and what these differences may imply when lexical richness is measured. Next, we present a new method for analyzing L2 learners' lexical profiles in oral production data, giving a detailed technical description of the creation of the tool. We then discuss pros and cons with frequency-based measures in general and present our solutions to some of the problems brought up. Finally, we suggest some potential areas of use and discuss some possible improvements of the method.

1. Background: a survey of methods designed to measure lexical richness in L2 production

In the study of L2 vocabulary, lexical richness can be seen as an umbrella term, covering four different dimensions: *lexical density, lexical diversity, lexical sophistication* and *proportion of errors* among the words used by an L2 learner (Read, 2000, pp. 200-201). Lexical density can be measured as the proportion of semantically full words (or lexical words) as opposed to function words. Lexical diversity, or variation, can be measured by the simple type/token ratio (TTR). The TTR is a calculation of the number of types divided by the number of tokens in a text. The basic problem with TTR is its sensitivity to text length, as is well known. As explained by McCarthy and Jarvis (2007, p. 460), "the more words (tokens) a text has, the less likely it is that new words (types) will occur". If a text is so long that certain words start to be repeated, high-frequency words

will be repeated more often as compared to low-frequency words, and this tendency will increase the longer the text is. Several measures have been proposed in order to solve the problem with text length. One example is the index of Guiraud (Guiraud, 1954), which is a type/token based measure that is supposed to be independent of text length. The index of Guiraud results from dividing the number of types by the square root of the number of tokens. For a long text, this procedure will result in a higher lexical richness than what would have been obtained with a simple TTR. However, according to Daller, Van Hout and Treffers-Daller (2003, p. 200) neither TTR nor the index of Guiraud are valid measures of lexical richness at later stages of L2 acquisition. A development of the Guiraud index is the advanced Guiraud, which takes in frequency as a factor (Daller et al., 2003). Furthermore, Malvern, Richards, Chipere and Durán (2004) have suggested the D measure, which is freely available in CHILDES. D models the falling TTR curve by calculating TTRs for samples of different text lengths, ranging from samples of 35 words to samples of 50 words, which are taken randomly from the text. However, in their critical evaluation of D, McCarthy and Jarvis (2007) conclude that even though the D measure was the most reliable of those investigated, it still retains a certain degree of sensitivity to text length.

Lexical sophistication is defined as the percentage of sophisticated or advanced words in a text. There are, however, different definitions of sophisticated/advanced vocabulary. Low-frequency words, for instance, are generally considered to be advanced and sophisticated (Laufer & Nation, 1995; Vermeer, 2004). It has even been suggested that words are learned in rough order of frequency (Cobb & Horst, 2004; Vermeer, 2004). The difficulty of words, as measured by their frequency, should therefore be taken into account when measuring the lexical richness of L2 learners. A method which relies on the raw frequency of words in the target language is the *Lexical Frequency Profile*, LFP (Laufer & Nation, 1995). The LFP measures the proportion of high-frequency words vs. the proportion of low-frequency words in a written text. All the words are divided into different categories, which have been established on the basis of frequency bands based on written language corpora (Laufer & Nation, 1995). *Vocabprofile* is a program that executes this categorization according to the following frequency bands: the 1000 most frequent word families, the next 1000 most frequent word families, and the Academic Wordlist, which contains the 570 most frequent word families drawn from academic texts (Coxhead, 2000, see also www.lextutor.ca/vocabprofile). The words that do not appear in any of these categories end up in the 'not-in-the-lists' category.[1]

1 There is also an updated version of Vocabprofile for English (but not for French), which distinguishes 20 different frequency bands.

Laufer and Nation (1995) have shown that the LFP measure is able to distinguish between different proficiency levels. The English version of LFP was validated by Laufer and Nation and there is also a French version, with the program *Vocabprofil*, also based on written data, which has been validated in a study of the oral production of advanced French L2 learners by Ovtcharov, Cobb and Halter (2006). It is interesting to note that Ovtcharov et al. actually used oral learner data and ran those against frequency bands based on written data. Still, they found significant differences between learners at different proficiency levels.

2. Lexical sophistication in written vs. spoken language

Even though Ovtcharov at al. (2006) were able to validate the French version of LFP using learners' oral production data, the appropriateness of comparing learners' *spoken* language with written data bases can be questioned. Lindqvist (2010) used the French version, Vocabprofil, comparing two groups at different proficiency levels.[2] In contrast to Ovtcharov et al. (2006), she found no significant differences between the two learner groups. She also conducted a qualitative analysis of the words classified in the not-in-the-lists category, and found that many words typical in oral French were classified in this category, such as *ben* ('well'), *ouais* ('yeah'), *rigolo* ('fun'), *prof* (short for 'teacher'), *sympa* ('nice'), although these are frequent in everyday speech. Lindqvist suggested that frequency lists based on L1 oral data should be used when investigating L2 learners' oral production. This has also been pointed out by Tidball and Treffers-Daller (2008, p. 311), who call for an oral version of the Vocabprofil program, so that oral data can be compared to an oral data base, which would better reflect the informants' lexical profile. For instance, the words *ben* and *ouais* are discourse markers that are often found in spoken language, but not in written production (McCarthy, 1998; Tidball & Treffers-Daller, 2008), so even if they are produced often by a learner a comparison to a written data base would give the impression that the learner uses rare words, and the conclusion that the learner in question has an advanced vocabulary might be wrong. According to McCarthy (1998, p. 122), frequency lists based on spoken language differ from those based on written sources. Generally, the differences between spoken and written language are considerable (see e.g. Linell, 2005, p. 28), something that must

2 The levels of proficiency of the learners were established on the basis of a morpho-syntactic analysis (cf. Barting & Schlyter, 2004).

have consequences at the lexical level of language. Considering this, there is a clear risk of running into validity problems when comparing spoken language to written corpora.

3. A new method for analyzing learners' lexical profiles in oral production data: the Lexical Oral Production Profile (LOPP)

Considering the background described above, and in order to avoid not only a written language bias (cf. Linell, 2005), but also methodological problems of validity, we set out to create a new tool for analyzing lexical sophistication in French and Italian L2, within the on-going project *Aspects of the advanced L2 learner's lexicon*.[3] We developed a lexical profiler explicitly for the analysis of spoken language. In order to create frequency bands based on spoken target language data, we used the *Corpaix* corpus for French and the *C-Oral-Rom* and *LIP* corpora for Italian.[4] We also developed a program that runs learner data against the frequency bands. In the following, we will describe the process of creating the tool.

3.1. SQL: a tool for manipulating data bases

SQL stands for *Structured Query Language* and is a declarative programming language initially developed at IBM with the purpose of manipulating big data bases. Work with data bases emerged in the 1960s due to cheaper storage and computing power (Wilton & Colby, 2005, p. 7), and the first scientific article discussing SQL was published in 1970 by the IBM researcher Codd (1970). SQL is now standardized by both the International Standards Organization (ISO) and by the American National Standards Institute (ANSI) (Jones et al., 2005, p. 2).

SQL is a data base management system allowing one to access and manipulate data bases. A data base could be described as a set of one or more tables organized in a systematic way or as "one or more large structured sets of persist-

3 This study is part of the research program High-Level Proficiency in Second Language Use, funded by the Bank of Sweden Tercentenary Foundation (grant M2005-0459).

4 An inherent problem with spoken language corpora is the relative limitations that the oral language mode implies, in terms of technical adjustments needed, transcription etc. As a consequence, these corpora are rather small, in comparison to, for example, the BNC.

ent data, usually associated with software to update and query the data" (The Free On-line Dictionary of Computing: http://foldoc.org/database). When working with sets of associated tables, i.e. retrieving, organizing, joining, counting and comparing table contents, work is very much facilitated if a query language such as SQL can be used.

3.2. Construction of the French and Italian frequency bands

The French frequency bands are based on the oral corpus Corpaix, compiled at the Université de Provence (Campione, Véronis, & Deulofeu, 2005). The corpus consists of about 1 million tokens based on different tasks such as interviews, conversations and meetings on different topics such as personal memories, travel, politics and professional experiences. A token-frequency list, based on Corpaix, has been created and published online at http://sites.univ-provence.fr/veronis/data/freq-oral.txt by Jean Véronis and that list was used when creating the French frequency bands discussed in the present study.[5] All tokens in the list were lemmatized with the software TreeTagger (Schmid, 1994, 1995) and later run through the software WordSmith (Scott, 2004) to calculate the frequency of each lemma. Hence, the final result consists of a lemma-frequency list composed of 2746 different lemmas.[6]

In regard to the Italian frequency bands, they were based on the already lemmatized versions of two different oral corpora: the LIP (Lessico di frequenza dell'italiano parlato) (De Mauro, Mancini, Vedovelli, & Voghera, 1993), which is freely available at the site BADIP (Schneider, 2008) and the C-Oral-Rom corpus (Cresti & Moneglia, 2005). The LIP corpus is based on several types of oral production: face-to-face conversations, telephone conversations, non-free dialogical interactions, monologues and radio and TV programs. C-Oral-Rom is based on both formal and informal speech, face-to-face conversations, telephone conversations and broadcasting. The social context of data collection is both private, within the family, and public, for example political speech and debate. A Perl programming language script was run on the XML versions of the two corpora in order to create a lemma-

5 Only tokens that appear ten times or more in the Corpaix corpus were added to the list created by Véronis.

6 This number has been corrected compared to earlier studies (Bardel, Gudmundson, & Lindqvist, 2012; Lindqvist, Bardel, & Gudmundson, 2011) in which the number of lemmas was estimated to 2766, due to a technical error. This small difference does not have any effect on the division of the lemmas into the frequency bands.

frequency list based on both LIP and C-Oral-Rom. The final result consists of a lemma-frequency list composed of 19962 different lemmas based on a total of 789070 tokens.

When creating the French and Italian frequency bands it was decided to use the lemma as counting unit instead of the word family, for the following reasons (for a more detailed discussion, see Lindqvist et al., 2011). A word family can include both derivations and inflected forms of a headword, which implies that the word family might include quite a high number of forms. For example, an Italian regular verb has six different forms in present tense: *canto, canti, canta, cantiamo, cantate, cantano* (from inf. *cantare*). This marking of person is compounded with marking of tense, aspect and modality (e.g. past tense of subjunctive 1st person plural: *cantassimo*). Hence, Italian has a very rich verb morphology. Furthermore a word family can also include nouns, adjectives, etc, whose relationships with the base are not always very transparent, such as *canzone* (song), *cantante* (singer) and, possibly, *cantautore* (a compound of *cantante* and *autore*, singer/songwriter). The fact that a learner uses one particular form does not necessarily mean that he or she has knowledge of all the related forms in the word family. This claim is particularly relevant in our research, which concerns oral production. It is plausible that the learner knows several word forms that are simply not used in one particular recorded session, which makes it impossible to draw any conclusions regarding how many forms related to a specific word family are actually known. Using the lemma as counting unit is an option that reduces the number of forms attached to a headword, even though this does not solve the problem completely. In conclusion, the French and Italian frequency bands described in this paper are different from the ones elaborated by Laufer and Nation (1995) and Cobb and Horst (2004), which are based on word families.

2746 lemmas from the French lemma-frequency list and 3127 lemmas from the Italian lemma-frequency list were divided into three frequency bands consisting of about 1000 lemmas each. Hence, band 1 includes the

Table 1. The French frequency bands

Band	Lemma range	Lemmas (n)	Tokens (n)	Relative token frequency (%)
1	1-986	986	896347	95.93
2	987-1939	953	28003	3.00
3	1940-2746	807	10034	1.07
Total		2746	934384	100

most frequent 1000 lemmas, band 2 the 2nd 1000 most frequent lemmas and band 3 the 3rd 1000 most frequent lemmas. The lemmas not appearing in any of these three bands are categorized as off-list lemmas, i.e. those not belonging to the most frequent 3000 lemmas in Italian or French. Table 1 shows the frequency distribution of the French frequency bands and table 2 the frequency distribution of the Italian frequency bands.

Table 2. The Italian frequency bands

Band	Lemma range	Lemmas (n)	Tokens (n)	Relative token frequency (%)
1	1-1019	1019	676098	91.82
2	1019-2047	1028	39726	5.39
3	2048-3127	1080	20526	2.79
Total		3127	736350	100

The tokens included in the French frequency bands (1-3) cover 93.44% of the total number of tokens included in the Corpaix corpus, and the tokens included in the Italian frequency bands (1-3) cover 93.32% of the total number of tokens included in the Italian corpus, i.e. the combination of LIP and C-Oral-Rom. As can be seen from the tables above, the number of lemmas included in the Italian frequency bands is slightly higher than that of the French bands. It can also be noted that the number of lemmas included in each band within each language varies between 807 and 986 for French and between 1019 and 1080 for Italian. The reason for this is that the line between two frequency bands must be drawn where two lemmas differ in frequency; for example, in the French list, all lemmas from rank 971 to 986 occur 50 times in the corpus, while the lemma ranked as number 987, *journal* (newspaper) occurs 49 times. *Journal* could not be included in the first frequency band since it would have been necessary to include all other lemmas that occur 49 times as well. The number of lemmas included in each band could therefore not be established and decided beforehand. The aim, however, was to distribute them as evenly as possible. It can be noted that more than 90% of all tokens that appear in the two corpora belong to band 1 and that only a small percentage belong to bands 2 and 3. The French and Italian frequency bands were imported into an SQL data base.

3.3. The lexical oral production profiler (LOPP): running analysis

French and Italian learner production can be compared to the frequency bands to measure the proportion of lemmas that fall within each frequency band. In order to do that, all data has to be lemmatized and information about lemma

frequency must be added. Other information, such as name of informant/name of recording, the language status (i.e. whether it's an L1 or an L2 speaker), and the linguistic level, i.e. proficiency level, can also be included.[7] Figure 1 shows part of an input file.

Figure 1. Part of a French input file

informant name	language status	lemma	lemma freq	linguistic level
lda4int	L2	le	74	6
lda4int	L2	être	71	6
lda4int	L2	on	70	6
lda4int	L2	avoir	58	6
lda4int	L2	je	58	6
lda4int	L2	de	51	6
lda4int	L2	un	44	6
lda4int	L2	que	41	6
lda4int	L2	oui	36	6
...	...	...	...	...

The following SQL query can be used to compare French learner data to the French frequency bands (named 'corpaixband').

```
(1)
SELECT
i.InformantName,
i.LinguisticLevel,
sum(LemmaFreq) as "number of lemmas",
sum(case when band = 1 then freq else 0 end) as "band 1",
sum(case when band = 2 then freq else 0 end) as "band 2",
sum(case when band = 3 then freq else 0 end) as "band 3",
sum(case when band is null then freq else 0 end) as "offlist"
FROM FrenchInputFile i
left outer join corpaixband b on i.lemma = b.lemma
group by InformantName
order by LinguisticLevel
```

In example (1) above, the content of the field/column 'LemmaFreq' from the table 'FrenchInputFile' is compared to that of 'corpaixband', creating an output file with information about the number of lemmas in the 'FrenchInputFile' belonging to band 1, band 2, band 3 and offlist. The result is grouped and ordered by 'InformantName' and 'LinguisticLevel' as shown in the figure below.

7 Proficiency level was operationalized as a 1-6 scale based on Bartning & Schlyter's (2004) framework, where 6 corresponds to a very advanced level.

Figure 2. Part of a French output file

informant name	linguistic level	number of lemmas	band1	band2	band3	offlist
Yvonne1int	4	1841	1573	14	7	247
Christina1int	4	1430	1190	15	14	211
Christina4int	4	1384	1159	30	10	185
Eva1int	4	1669	1385	21	9	254
Eva4int	4	1430	1199	12	5	214
Malena1int	4	2068	1737	16	6	309
Mona1int	4	1249	988	14	3	244
Pernilla1int	4	1126	976	14	8	128
Pernilla4int	4	1171	963	11	3	194
Yvonne4int	4	2284	1904	23	6	351
Ida4int	6	1488	1224	29	9	226
Kerstin1int	6	1764	1457	22	4	281

The output shown in figure 2 can easily be exported to an Excel spreadsheet where the number of lemmas can be converted into proportions. The following figures illustrate the lexical frequency profile, in terms of number and proportions of lemmas, for Eva4int.

Figure 3. Lexical richness output: number of lemmas in Eva4int

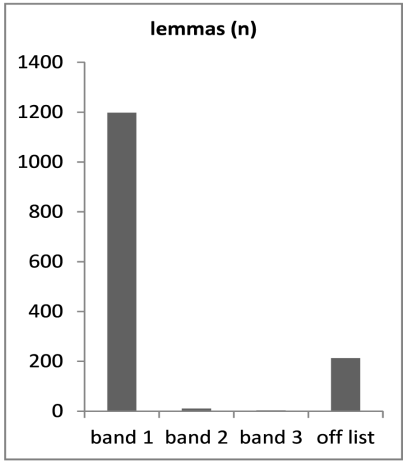
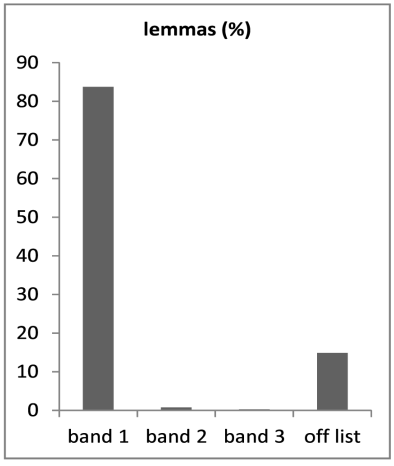


Figure 4. Lexical richness output: proportion of lemmas in Eva4int



Another useful query provides information about the informant's name, the lemma, the frequency of the lemma, the linguistic level of the informant, and the band to which the lemma belongs. The query is shown in example (2) and it returns an output file represented in figure 5.

```
(2)
select
i.InformantName,
i.lemma,
i.LemmaFreq,
i.LinguisticLevel,
b.band
from FrenchInputFile i
left outer join corpaixband b on i.lemma = b.lemma
```

Figure 5. Part of a French output file

informant name	linguistic level	lemma	lemma freq	band
lda4int	6	le	74	1
lda4int	6	être	71	1
lda4int	6	penser	5	1
lda4int	6	valise	4	offlist
lda4int	6	aimer	3	1
lda4int	6	arabe	3	1
lda4int	6	avion	3	2
lda4int	6	vouloir	3	1
lda4int	6	aéroport	2	offlist
lda4int	6	alors	2	1
lda4int	6	beaucoup	2	1
lda4int	6	bon	2	1
lda4int	6	brancher	2	2
lda4int	6	cinquante	2	1
lda4int	6	comprendre	2	1
lda4int	6	contenter	2	2
lda4int	6	devant	2	1
lda4int	6	inscrire	1	3
lda4int	6	intéressant	1	1
lda4int	6	irréel	1	offlist
...	...	...	...	...

As can be seen from the output file in figure 5, the last column indicates the band to which the lemma belongs. This is useful information when single lemmas have to be studied and analyzed.

4. Pros and cons with frequency-based measures

Two important advantages with the lexical frequency profiling analysis are that it is able to distinguish between proficiency levels in oral production and that

this measure of lexical richness seems to correlate with the other measures of proficiency used in our earlier studies. However, there are also some important drawbacks with this kind of measure in general. Some of them will be discussed at the end of this paper. There are also problems related to the frequency criterion *per se*. The method relies exclusively on (low-) frequency as a criterion of high level proficiency (or difficulty for the learner). Other factors that may have an impact on learnability (and lexical richness) are cognateness and the role of teaching materials (cf. Horst & Collins, 2006; Milton, 2007). Horst and Collins showed that the use of cognates decreased with higher proficiency, suggesting that cognates (although of low frequency) are not indicative of an advanced vocabulary, in the sense of LFP. As for the role of teaching materials, Milton has pointed out that words that are introduced early, covering certain thematic fields, like travelling or eating out, are learned early, even though they are not used in everyday speech by native speakers, and these words are erroneously classified when regarded as advanced vocabulary. These issues were explored in Bardel and Lindqvist (2011), which led to certain modifications of the LOPP method. These modifications are described in the following section.

4.1. LOPPa: further elaborations of LOPP

Bardel and Lindqvist (2011) investigated the role of cognates and thematic vocabulary in two learners of French and two learners of Italian at different proficiency levels, focusing on the use of low-frequency words. They found that among the low-frequency words produced by the learners there were many cognates and thematic words related to teaching materials, i.e. words, although infrequent, that could be considered rather easy for a Swedish learner of French or Italian. The authors therefore suggested an elaboration of the LOPP tool in order to measure lexical richness in a way that takes not only the proportion of words belonging to a certain frequency band into account, but also the cognate-factor and the thematic word-factor. A new tool, LOPPa, was therefore created. While the old tool, henceforth LOPPf, splits the learner data into three frequency bands, LOPPa classifies each word in the learner data as either basic or advanced.[8] The basic vocabulary is composed of a combination of high frequency words, basic cognates and basic thematic words, while the advanced vocabulary is composed of low-frequency words, advanced cognates and advanced thematic words. In order to operationalize the concept of basic cognates and basic thematic words vs. advanced cognates and advanced thematic words, teachers' judgements were used (cf. Tidball & Treffers-Daller, 2008). A full description

8 *a* stands for advanced and *f* for frequency.

of the methodology used to carry out the teachers' judgement test can be found in Bardel et al. (2012).

In order to evaluate the LOPPa tool, data from a previous study carried out with the LOPPf tool (Lindqvist et al., 2011) were re-analyzed with the LOPPa tool (Bardel et al., 2012). It was found that the distinction between basic and advanced words resulted in a higher intra-group homogeneity compared to the purely frequency based perspective. Thus, by taking cognateness and the notion of thematic words into consideration, the lexical richness measure improved, an improvement that was shown by an increased effect size as expressed by eta^2.

5. Potential areas of use of the method

On the basis of our research we can claim that there are two main advantages with lexical frequency profiling analyses: (1) They are able to distinguish between proficiency levels in oral production. This has been shown both for the method relying only on frequency (Lindqvist et al., 2011) and for the elaborated version of the method, which takes cognates and thematic vocabulary into account (Bardel et al., 2012). (2) LOPPa provides results that seem to correlate with other measures of proficiency used in our earlier studies (mainly measures of morpho-syntactic development).

Another advantage that we would like to point out is that it is possible to conduct both quantitative and qualitative analyses using LOPPa, as opposed to using formulas of lexical richness, e.g. D or TTR. The procedure of LOPPa is to first provide a quantitative result, i.e. the division of the lemmas into bands. In a second phase, it is possible to make an in-depth analysis of the words actually used, by looking at the lists provided by the program. This is possible for a whole data set as well as for individual learners. By making such a thorough analysis it is also possible to continuously improve the method by analyzing the words that appear in the off-list for instance. It is plausible that new cognates and words belonging to thematic vocabulary will appear in the off-list when new data is used in the program. We also believe that the method could be used for pedagogical purposes, for example in order to assess learners' lexical richness in oral production. Teachers could use the basic/advanced word list as a point of reference in vocabulary teaching. The method is also suitable for self-assessment, if learners are given the possibility to analyze their own production within a specific course component at higher levels of education.

It has to be admitted that there are some limitations to the method at this stage of our research. One of the limitations concerns the fact that it is oriented towards learners with Swedish as their L1 and French or Italian as their L2 (and also taking into account that English is an additional second language for all

learners). This certainly limits the number of potential users. However, given the detailed description of the elaboration of the method provided in this paper, there are good possibilities to adapt it for use with other languages. Another limitation is that the method is most suitable for oral data. As we have discussed elsewhere, it is preferable to compare learner data to the same type of data in the target language, as word frequency may differ between oral and written language.

There are also some important drawbacks with this kind of measure of lexical richness in general. One is that it only taps formal aspects of word knowledge. Deep knowledge of vocabulary is not accounted for, e.g. use of words with multiple meanings or use of multi-word units (cf. Nation, 2006; Cobb, this volume). Furthermore, another aspect that remains ignored is non-targetlike use of target language forms. Possible solutions to these problems will be discussed in the following section.

6. Possible improvements of LOPPa

There are several aspects that must be learned in order to achieve complete knowledge of a word: form (spoken and written, i.e. pronunciation and spelling), word structure (morphology), syntactic pattern of the word in a phrase and sentence, meaning (referential – including multiplicity of meaning and metaphorical extensions of meaning; affective – the connotation of the word; pragmatic – the suitability of the word in a particular situation), lexical relations of the word with other words (e.g. synonymy, antonymy, hyponomy) and collocations. All these aspects can be more or less well known. The more advanced a learner, the more aspects of a word are likely to be known, and the more developed are the different aspects, for example, more meanings of a homograph are known, more synonyms, more collocations and idiomatic expressions are mastered (Laufer, 1997, p.141).

Qualitative knowledge about the single word is sometimes referred to as *depth*. In his attempt to pinpoint what researchers have in mind when investigating depth of knowledge, Read (2004) distinguishes three approaches to vocabulary learning in the literature, *comprehensive word knowledge, precision of meaning* and *network knowledge.* According to the first approach, depth covers different types of knowledge of a word, like those indicated by Laufer (1997, p. 141), all of which, if they are fulfilled, can be called *comprehensive word knowledge*. With *precision of meaning*, Read (2004, p. 211) refers to "the difference between having a limited, vague idea of what a word means and having much more elaborated and specific knowledge of its meaning". It seems problematic to establish a criterion for precise knowledge. Typically, the criterion is that of the adult native speaker. However, as Read (2004, p. 213) points out, "knowl-

edge of specialized, low-frequency vocabulary reflects in the first instance a person's level and field of education but also their social background, occupation, personal interests and so on". Depth can also be understood as *network knowledge*, i.e. the incorporation of a word into the network surrounding it in the mental lexicon. Word knowledge is sometimes thought of as a network, and words as interconnected nodes. The nodes are interconnected in different dimensions, thematically, phonologically, morphologically, conceptually etc. (Vermeer, 2001, p. 218; Meara, 2009; Gyllstad, this volume).

Two aspects of deep knowledge that are crucial parts of complete word knowledge concern the multiple meaning of polysemic words or homographs and the meaning of multi-word units. Knowing several meanings of a single word form is a kind of deep knowledge that is referred to as *range of meaning* in addition to precision of meaning (see above) by Read (2000, p. 92). The role of context is essential for the interpretation of the meaning of words, and this becomes obvious when dealing with words with multiple meanings and with multi-word units. In lexical frequency profiling, these two aspects become problematic, since the profilers normally do not take context into account. A disadvantage with frequency-based measures such as LFP or LOPPa is that they do not account for the frequency of each meaning attached to a word form (see also Nation, 2006, p. 66). A homograph like French *louer* will always be categorized in the same frequency band independently of the meaning attached to it (*rent* or *praise*), even though the different meanings of the word may not be equally frequent (see Cobb, this volume). It has been suggested that more advanced learners know more meanings of a word than less advanced learners (cf. Bensoussan & Laufer, 1984). It would therefore be a great advantage if lexical profilers could be adapted in order to account for the frequency of the meaning of the word used in a particular context. In that way, the measure would be sensitive to the possible variation of frequency of different meanings of words in the learners' input.

Another qualitative aspect of word knowledge is the knowledge and ability to use multi-word units. A multi-word unit can be defined as a particular combination of words that generates one meaning (see Henriksen, this volume, for an overview of different definitions). One approach to multi-word units is that of Wray (2002), according to whom such combinations of words seem to be retrieved as a whole unit from memory (Wray, 2002, p. 9). This usage of particular word combinations cannot be measured in the LFP, nor in LOPPa, because the programs use graphic criteria to define a word. This means that expressions in French like *tout le monde* (everybody) or *tout à fait* (exactly) will be regarded as three separate words and not as one unit that generates one meaning. Moreover, the words contained in a multi-word unit may belong to different frequency bands. As for *tout à fait*, *tout* and *à* belong to Band 1, while

fait is an off-list word. Treating these words separately means that the number of words categorized as highly frequent will rise, although this may not correspond to the frequency of the whole expression in the target language input. In order to account for the frequency of multi-word units, we would have to find a way to integrate them in the frequency lists. It is encouraging to see that work in this direction has started for English (Cobb, this volume; Martinez & Schmitt, 2012). However, considering our approach in the LOPPa framework, we find it pertinent to include multi-word units that are cognates (Wolter & Gyllstad, 2011) and thematic in a basic and an advanced vocabulary.

How could this be accomplished within the LOPPa framework? Every multi-word unit present in the corpus to be analyzed must be tagged as a unit in order to make it appear as a unit and not as several different words. This would lead to a non-match with the baseline corpora, if they are not tagged in exactly the same way, and consequently the multi-word units would end up in the off-list among the low-frequent advanced words. If the aim is to get a picture of the role of frequency for vocabulary learning, as in the LFP, one must make them appear in the frequency bands they actually belong to, and in order to do this the actual frequency of the multi-word units must be looked up in the corpora used as baseline data. Of course, the same goes for the multiple meanings of words. Words occurring in the baseline corpora must be sorted into frequency bands on the basis of the meaning they have in context.

Another important aspect, which is not accounted for in lexical profiling analyses, is the use of words that do not exist in the TL. In fact, non target-like word forms and non target-like use of words (although correct at the formal level) represent an important aspect of vocabulary knowledge. Our main focus thus far has been on the vocabulary use by relatively advanced learners, but earlier research has shown that cross-linguistic influence occurs more frequently at the earlier stages of development (Lindqvist, 2009; Williams & Hammarberg, 2009 [1998]). It is important to integrate this aspect when analyzing the lexical profile of learners. Moreover, as noted above, Read (2000) considers that the proportion of errors is one aspect of lexical richness.

Non target-like use can be instances of code-switching, lexical inventions or other deviant forms of words in the TL (Bardel & Lindqvist, 2007; Dewaele, 1998; Williams & Hammarberg, 2009 [1998]). Vocabprofile gives the instruction to remove code-switches and other deviant forms, and this was also done in the Laufer and Nation (1995) study. We followed this methodology in the LOPPf/a analyses. The main reason for that is that if they had been kept, words belonging to another language than the TL would end up in the off-list, thus adding to the proportion of advanced words. However, in our view, code-switches are also part of the learner's vocabulary, and have something to say about the level of vocabulary proficiency. Moreover, the fact that a learner uses

a correct TL word form does not automatically imply that it is appropriate in the context. However, since lexical profiling methods are not sensitive to context, this type of deviance will not be captured. An example of a word (in this case a multi-word unit) from one of the learners in the present study is the expression *tout le monde* (everybody), which is used in the sense of *le monde entier* (the whole world). The non target-like use of the expression cannot be captured without a closer look at the context.

7. Conclusions

As we have shown, several efforts have been made within the project *Aspects of the advanced L2 learner's lexicon*, to create and improve a tool for lexical profiling of Swedish L2 learners' oral production of French and Italian. In a number of steps we have improved our original method LOPP, but there are still many things to develop further. On top of the ideas put forward in this chapter, given that the method is now only available to the research group, an important step forward would be to make the method and the data accessible to other users by providing a user-friendly interface.

References

Bardel, C., Gudmundson, A., & Lindqvist, C. (2012). Aspects of lexical sophistication in advanced learners' oral production: Vocabulary acquisition and use in L2 French and Italian. *Studies in Second Language Acquisition, 34*(2), 269-290.

Bardel, C. & Lindqvist, C. (2007). The role of proficiency and psychotypology in cross-linguistic influence. A study of a multilingual learner of Italian L3. In M. Chini, P. Desideri, M.E. Favilla & G. Pallotti (Eds.), *Atti del XI congresso internazionale dell'Associazione italiana di linguistica applicata. Napoli 9-10 febbraio 2006* (pp. 123-145). Perugia: Guerra.

Bardel, C. & Lindqvist, C. (2011). Developing a lexical profiler for spoken French and Ialian L2: The role of frequency, cognates and thematic vocabulary. In L. Roberts, G. Pallotti, & C. Bettoni (Eds.), *EUROSLA yearbook 11* (pp. 75-93). Amstedam: Benjamins.

Bartning, I. & Schlyter, S. (2004). Itinéraires acquisitionnels et stades de développement en français L2. *Journal of French Language Studies, 14*(3), 281-289.

Bensoussan, M. & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading, 7*(1), 15-32.

Campione, E., Véronis, J., & Deulofeu, J. (2005). The French corpus. In E. Cresti, & M. Moneglia (Eds.), *C-ORAL-ROM: Integrated reference corpora for spoken romance languages* (pp. 111-133). Amsterdam: Benjamins.

Cobb, T. & Horst, M. (2004). Is there room for an academic wordlist in French? In P. Boogards, & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 15-38). Amsterdam: Benjamins.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM, 13*(6), 377-387.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Cresti, E. & Moneglia, M. (2005). *C-ORAL-ROM: Integrated reference corpora for spoken romance languages*. Amsterdam: Benjamins.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics, 24*(2), 197-222.

De Mauro, T., Mancini, F., Vedovelli, M., & Voghera, M. (1993). *Lessico di frequenza dell'italiano parlato* (1st ed.). Milano: Etaslibri.

Dewaele, J. (1998). Lexical inventions: French interlanguage as L2 versus L3. *Applied Linguistics, 19*(4), 471-490.

Guiraud, P. (1954). *Les caractéristiques statistiques du vocabulaire*. Paris: Presses Universitaires de France.

Horst, M. & Collins, L. (2006). From *faible* to strong: How does their vocabulary grow? *Canadian Modern Language Review, 63*(1), 83-106.

Jones, A., Stephens, R., Plew, R. R., Garrett, B., & Kriegel, A. (2005). *SQL functions programmer's reference (programmer to programmer)*. Indianapolis: Wiley Pub.

Laufer, B. (1997). The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. In J. Coady & T. N. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 20-34). Cambridge: Cambridge University Press.

Laufer, B. & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-322.

Lindqvist, C. (2009). The use of the L1 and the L2 in French L3: Examining cross-linguistic lexemes in multilingual learners' oral production. *International Journal of Multilingualism, 6*(3), 281-297.

Lindqvist, C. (2010). La richesse lexicale dans la production orale de l'apprenant avancé de français. *Canadian Modern Language Review, 66*(3), 393-420.

Lindqvist, C., Bardel, C., & Gudmundson, A. (2011). Lexical richness in the advanced learner's oral production of French and Italian L2. *IRAL, 49*(3), 221-240.

Linell, P. (2005). *The written language bias in linguistics*. London: Routledge.

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke: Palgrave Macmillan.

Martinez, R. & Schmitt, N. (2012). A phrasal expression list. *Applied Linguistics, 33*(3), 299-320.

McCarthy, M. (1998). *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.

McCarthy, P. M. & Jarvis, S. (2007). *Vocd*: A theoretical and empirical evaluation. *Language Testing, 24*(4), 459-488.

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition*. Amsterdam: Benjamins.

Milton, J. (2007). Lexical profiles, learning styles and the construct validity of lexical size tests. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 47-58). Cambridge: Cambridge University Press.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review 63*(1), 59-82.

Ovtcharov, V., Cobb, T., & Halter, R. (2006). La richesse lexicale des productions orales: Mesure fiable du niveau de compétence langagière. *The Canadian Modern Language Review, 61*(1), 107-125.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Read, J. (2004). Reserch in teaching vocabulary. *Annual Review of Applied Linguistics, 24*, 146-161.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *International Conference on New Methods in Language Processing,* Manchester, UK.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. *Proceedings of the ACL SIGDAT-Workshop,* Dublin, Ireland. 1-9.

Schneider, S. (2008). *BADIP.* Retrieved 10/10, 2008, from http://languageserver.uni-graz.at/badip/badip/home.php

Scott, M. (2004). *WordSmith tools version 4*. Oxford: Oxford University Press.

Tidball, F., & Treffers-Daller, J. (2008). Analysing lexical richness in French learner language: What frequency lists and teacher judgment can tell us about basic and advanced words. *French Language Studies, 18*(3), 299-313.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*(2), 217-234.

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Boogards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 173-189). Amsterdam: Benjamins.

Williams, S. & Hammarberg, B. (2009 [1998]). Language switches in L3 production: Implications for a polyglot speaking model. In B. Hammarberg (Ed.), *Third language acquisition* (pp. 28-73). Edinburgh: Edinburgh University Press.

Wilton, P. & Colby, J. W. (2005). *Beginning SQL*. Indianapolis: Wiley.

Wolter, B. & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics*, *32*(4), 430-449.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.