Corpus-based versus intuition-based lexicography: defining a word list for a French learners' dictionary.

Serge Verlinde and Thierry Selva
Modern Language Institute, K.U. Leuven (Belgium)

1. Introduction

Although French lexicographers were among the first to integrate corpus-analysis into the dictionary-making process, with the *Trésor de la langue française* project in the early seventies and its corpus of 170 million words, corpus-based lexicography is certainly not a common practice in contemporary lexicography in France. There are a few exceptions, however, e.g. the *Oxford-Hachette English-French/French-English* translation dictionary (Corréard 1997) and the *Dictionnaire d'apprentissage du français des affaires* (DAFA - Binon, Verlinde, Van Dyck, Bertels 2000), but mainstream lexicography is undoubtedly intuition-based.

As far as we know, no comparative studies have been made of the results of the two lexicographic approaches. The aim of this paper is to present such a comparative study on a selection of words to be described in a learners' dictionary of French. On the one hand, we have a recent learners' dictionary, which has been published by one of the leading French dictionary publishers (Dictionnaires Le Robert), where the selection of the entries is intuition-based (*Dictionnaire du français* - DF, Rey-Debove 1999). On the other hand, we have the DAFLES (*Dictionnaire d'apprentissage du français langue étrangère ou seconde*, an (electronic) French learners' dictionary we are presently working on.

We try to use an objective frequency criterion to select the words and multiword units described in our dictionary. Therefore we use an automated statistical analysis of a 50 million word corpus of newspaper texts, taken from the 1998 issues of *Le Monde* (France) and *Le Soir* (French speaking part of Belgium).

In the two first sections of this paper, we will present this corpus and the analyses made on it. In the third section, we will compare the two lists of words in order to reveal the most important differences between these two lists. In the last section, we will make another comparison, namely with the only large frequency list existing for French, which has been published along the *Trésor de la langue française* (TLF), the *Dictionnaire des fréquences* (Imbs 1971).

2. The corpus

Two important questions arise when building a corpus: its representativeness and its size. For the French language there is currently no project like the British National Corpus (BNC 2000) or the Bank of English (BOE 2000); therefore we must rely on the texts that are freely accessible (see Verlinde, Selva (forthcoming) for an overview of available corpora for French). The choice is limited to literature in the public domain and newspaper texts published on archive CD-ROMs. In order to cover actual language, we have chosen the 1998 issues of two newspapers: *Le Monde* (France) and *Le Soir* (French speaking part of Belgium). Both CD-ROMs permit the texts of all articles to be exported. In the case of *Le Soir*, exporting can be done by date or by newspaper section. In the case of *Le Monde*, there is no clear-cut classification of the articles. Therefore, only exporting by date makes it possible to export all articles. The corpus has a total size of 54 260 926 words, both subcorpora having approximately the same size.

We cannot say that our corpus is perfectly balanced, but it is made up of the kind of texts that the potential users of our dictionary will have to deal with.

At the first stage, we cut all documentary information about the articles from the corpus. Indications about source, date and page have been coded separately in the format of the text analysis software we use (see below). The whole corpus was then tagged and lemmatised with the Cordial-software, splitting up all multi-word units like *chemin de fer* and *pomme de terre* and removing proper nouns. The result of this analysis was processed in order to restore the aspect of the original texts. We submitted the entire lemmatised corpus (51 845 143 words) to Wordcruncher, a well-known text analysis tool. As Wordcruncher was not able to merge both subcorpora, we have merged the two separate frequency lists of both subcorpora to create a frequency list for the whole corpus. This frequency list has been corrected on some minor points. For example, frequent words written with a

hyphen that were split up during the lemmatisation process have been extracted from the original corpus and added to the list. Some errors of lemmatisation have also been corrected.

Our corpus is smaller than the two big English corpora but it seems to be large enough, taking into account our objective of writing a learners' dictionary with a selection of the most common vocabulary, collocations and grammatical structures of the current language.

3. Frequency list and dictionary word list

For the lexicographer, it is particularly difficult to define the importance of a dictionary word list. We see that in many cases, the number of entries (macrostructure of the dictionary) is much more important than the content of each entry (microstructure of the dictionary). Thus, the accent lies on single words more than on word combinations (collocations), for instance. This is paradoxical because, for productive purposes, learners need this information on collocations much more (Bogaards 1996, 1998) than an impressive number of isolated words, many of which will never be looked up. We decided provisionally to limit the word list to 12 156 words, selecting all lemmas that appear at least 100 times in our corpus. These 12 156 lemmas represent approximately 93.14% of all the words of the corpus, proper nouns not included. Extending this word list to 22 000 words, as in the DF, would only increase the coverage of the texts by 1%.

It is surprising to see that this limited list contains a large number of words that are very common in spoken language: *maman*, *papa*, *job*, *sympa*, *bosser* for example. There are also a lot of words that should perhaps not appear in a learners' dictionary because they are immediately linked to current affairs (*bosniaque*, *kosovar*) or to the local pages of the newspapers (*brabançon*, *brainois*, *borain*).

Another aspect of vocabulary use that can be easily studied with the frequency list is the generalization of English words in French. It is well known that French authorities follow a policy of "defending" the French language by suggesting, quite systematically, French equivalents for English words. We might suppose that newspapers, mainly the French ones, wouldtry to reinforce this policy, but this seems not to be the case in light of the frequency of some English words (table 1).

	frequency *Le Monde*	frequency *Le Soir*
business	446	471
coach	83	1424
cool	108	169
design	305	312
fast-food	42	88
goal	69	108
holding	573	505
joint(-)venture	84	95
leasing	29	69
lobbying	137	114
marketing	847	767
team	76	590
trader	48	35
Web/web	1057	514

Table 1: frequency of some English words in the *Le Monde* and *Le Soir* corpora

Even for those words that have an accepted and well-know French equivalent (*affaires* for *business*, *but* for *goal*, *équipe* for *team*), the English word seems to be used quite regularly. In some cases, the French equivalent is not used at all (*mercatique* for *marketing*).

The fact that we are working with corpora from two different language communities makes it also possible to compare the vocabulary used in both communities and to extract words that are specific to one of these communities by comparing the relative frequency of their occurrences in both corpora. Table 2 shows an extract of the list of typical French and Belgian words and abbreviations.

typical French words	typical Belgian words
ballottage	Échevin
préfectoral	maï eur/mayeur
départemental	Communal
baccalauréat	Deputation
minitel	Tram
préfet	Subside
lycéen	play-off
intéressement	Coach
cantonal	Voirie
interministériel	Urbanistique
typical French abbreviations	typical Belgian abbreviations
mdc	Asbl
insee	Prl
cfdt	Cpas
smic	Psc
cgt	Rtbf

Table 2: list of typical French and Belgian words and abbreviations

Such information on geographical variants is rarely mentioned in the essentially France-oriented French dictionaries.

3. Corpus frequency list and word list of the DF

The DF is in fact the first learners' dictionary of French for twenty years. The objective, presenting the words of both everyday conversation and the press (Rey-Debove 1999: VII), is very close to our objective, and to the objective of every learners' dictionary in general. As the authors do not say that they integrated a corpus analysis, it is possible to make a comparison between a corpus-based approach and an intuition-based approach, at least for the word list of the dictionary. Similar comparisons could be made for the collocations and the syntactic structures that are described in the dictionary. Table 3 shows to what extent the word list of the DF matches the words of the corpus frequency list.

corpus frequency ranges	number of words not mentioned in the DF	percent	cumulative frequency	cumulative percent
0-500	0	0	0	0
501-1000	2	0,4	2	0.2
1001-1500	3	0,6	5	0.3
1501-2000	1	0,2	6	0.3
2001-2500	10	2	16	0.6
2501-3000	16	3,2	32	1.1
3001-3500	18	3,6	50	1.4
3501-4000	28	5,6	78	2
4001-4500	40	8	118	2.6
4501-5000	45	9	163	3.3
5001-5500	48	9,6	211	3.8
5501-6000	61	12,2	272	4.5
6001-6500	58	11,6	330	5.1
6501-7000	67	13,4	397	5.7
7001-7500	39	7,8	436	5.8
7501-8000	87	17,4	523	6.5
8001-8500	80	16	603	7.1
8501-9000	102	20,4	705	7.8
9001-9500	120	24	825	8.7
9501-10000	115	23	940	9.4

10001-10500	110	22	1050	10
10501-11000	129	25,8	1179	10.7
11001-11500	154	30,8	1333	11.6
11501-12000	124	24,8	1457	12.1

Table 3: corpus frequency ranges and DF word list

The conclusion that can be drawn from this table is that 12.1% of the 12 000 most frequent words of our corpus do not appear in the DF. The differences in coverage are limited up to frequency 4000 with a difference less than 10%. From frequency 4 000 on, and mainly from frequency 8 500 on, the differences in coverage increase seriously (up to 20% and more).

In the list of 'forgotten' words and abbreviations, we find *investisseur*, *budgétaire*, *entité*, *concertation*, *restructuration*, *infrastructure*, *forum*, *info*, *privatisation*, *amendement* for example. These words need to be mentioned in a general purpose dictionary.

When we have a look at the words mentioned in the DF that do not appear in our frequency list, we notice that these words can not really be considered as current words (table 4, excerpt from the beginning of the letter A).

a fortiori	abêtissant	abreuvoir	accessoiriste
à gogo	abjurer	abricotier	accotement
à jeun	ablution	abrutir	accouder (s')
a.z.t.	aboiement	abrutissant	accoudoir
abasourdi	abois (aux)	abscisse	accoutrement
abat-jour	abominablement	absenter (s')	accoutrer
abats	abortif	abyssin	accroupir (s')
abattant	aboutissants	acadien	accumulateur
abattis	abracadabrant	acariâtre	accus
abêtir	abrasif	accablement	achalandé

Table 4: DF words not appearing in the corpus frequency list

The authors of the DF identify the frequent and important words by marking them with a blue triangle. In our learners' dictionary we classify the words into six frequency ranges (table 5).

frequency range	Range	occurrences	text coverage
1	<= 427	>= 11 183	66 %
2	<= 990	>= 5 273	75 %
3	<= 1 926	>= 2 482	82 %
4	<= 3 920	>= 854	88 %
5	<= 12 156	>= 100	93 %
6		< 100	100 %

Table 5: DAFLES frequency ranges

Both frequency indications can be linked and compared. Once again, the intuitive approach seems to be less rigorous than a corpus-based approach: words like *acajou*, *adipeux*, *ablation*, *affairé* and *affublé* are "frequent and important" according to the authors of the DF but not *à*, *année*, *américain*, *allemand*, *afin de/que*.

Looking into detail at the whole list of "frequent and important" words for the letter A of the DF reveals however also some weaknesses of our corpus-based approach. Some everyday life words as *s'absenter* do not appear into our frequency list. They need certainly to be added to a learners' dictionary word list.

4. Comparing two corpus-based frequency lists: literature and newspapers

As it was mentioned above, the *Dictionnaire des fréquences* (Imbs 1971) is a frequency list published along the TLF. It is the only frequency list based on a large corpus (170 million words) for

French with literary texts from the beginning of the nineteenth century to the sixties. We selected the first 12 174 lemmas of this list in order to compare them to our frequency list.

It is not surprising that a lot of current words as *régional*, *match*, *euro*, *championnat*, *football*, *culturel*, *télévision* and *festival*, for example, do not appear in the frequency list of the *Dictionnaire des fréquences*. Words that do not appear in our list mainly characterise personal feelings (*sottise*, *fâché*, *gémir*, *tressaillir*) and things that do not exist anymore (*pardessus*, *sou*, *écu*).

In addition to everyday life words mentioned above, words expressing feelings in general form a second important group of words to be added to the word list of our dictionary.

5. Conclusion

From the comparison of both lexicographic approaches (corpus-based and intuition-based), we can conclude that corpus-based lexicography gives a strong and necessary empirical evidence to the lexicographer's personal intuition, even if this personal intuition remains helpful in filling the gaps in our corpus.

These gaps are undoubtedly due to the fact that the corpus is unbalanced. Taking into account this observation, there is a strong need to design and construct for French, and for other languages as well, a carefully selected corpus with a large variety of texts, in order to improve the quality of (learners') dictionaries, and vocabulary learning and teaching in general.

References

Binon J, Verlinde S, Van Dyck, J, Bertels A 2000 *Dictionnaire d'apprentissage du fraçais des affaires*. Paris, Didier. (an electronic version of this dictionary can be found at http://www.projetdafa.net).

Bogaards P 1996 Dictionaries for learners of English. *International Journal of Lexicography* 9(4): 277-320.

Bogaards P 1998 Des dictionnaires au service de l'apprentissage du français langue étrangère. *Cahiers de lexicologie* 72(1): 127-167.

Corréard M-H (ed.) 1997² *The Oxford-Hachette French Dictionary*. Paris-Oxford, Hachette-OUP.

Imbs P 1971 *Dictionnaire des fréquences. Vocabulaire littéraire des XIXe et XXe siècles, I - Table alphabétique, II - Table des fréquences décroissantes*. Nancy-Paris, CNRS-Didier.

Rey-Debove J (ed.) 1999 *Dictionnaire du français. Référence. Apprentissage*. Paris, CLE International-Dictionnaires le Robert.

TLF. Imbs P 1971-1994 *Trésor de la langue française*. Paris, CNRS-Gallimard.

Verlinde S, Selva Th forthcoming Nomenclature de dictionnaire et analyse de corpus. *Cahiers de lexicologie*.

Websites

BNC 2000: http://info.ox.ac.uk/bnc/

BOE 2000: http://titania.cobuild.collins.co.uk/boe_info.html