Monitoring Dictionary Use in the Electronic Age

Serge Verlinde and Jean Binon
Leuven Language Institute, K.U.Leuven, Belgium

The way in which a user consults a dictionary, navigates through a dictionary article and finds an answer to specific questions is a popular area of research in metalexicography. The successful development of online dictionaries opens new prospects in this area of research. Log files of online dictionaries may provide interesting 'free implicit feedback' (de Schryver and Joffe 2004: 187). Thanks to its task- and problem-oriented interface, the Base lexicale du français (BLF) allows us to track all dictionary users' actions in a natural setting, outside any controlled research environment. Using these data, it should be possible to make well justified decisions on dictionary design.

1. Monitoring dictionary use...

1.1. ...in the paper age

Welker (2006) gives an interesting overview of the wealth of literature available on this topic: he lists more than 200 studies. In spite of all this research effort, Bogaards (2003: 26-33) concludes that 'uses and users of dictionaries remain for the moment relatively unknown'. Data is often too vague, conclusions very general.

Nevertheless, according to Bogaards (2003: 28), if we want to bring the dictionary closer to the user, it is important to take further steps in the study of user behaviour.

1.2. ...in the electronic age

De Schryver and Joffe (2004: 187) reveal that a tracing and logging system, integrated into an electronic dictionary, has long since been suggested as a way to improve dictionaries. However, data recorded by most tracing and logging systems are also often relatively superficial, counting the number of user accesses per hour/day/month, the total number of accesses per country, the identification of the visiting server, visit duration, etc. Such statistics may be provided by *Awstats*, which is a free software distributed under the GNU General Public License, *Google Analytics* or by commercial products, like *Logaholic*.

However, the lists of words or word combinations submitted are much more interesting from a lexicographical point of view (see for example the list on the site of the Cambridge Advanced Learner's Dictionary). De Schryver and Joffe (2004) suggest using this information to enrich the dictionary content. This information could also be helpful in testing the success rate of search algorithms when they are faced with 'messy' user input (Měchura 2008: 1296).

More advanced statistics and data may be obtained by a specific software application suite for compiling dictionaries or terminology lists like *TshwaneLex*. De Schryver (2004: 191-192) illustrates how one can focus on individual users and analyze their particular searches.

However, even with all this data, we still do not know exactly how users read dictionary articles, i.e. skimming and scanning.

2. The Base lexicale du français

The *BLF*'s access structure is task- and problem-oriented and based on the users' needs, which leads to a limited number of occasional or more systematic consultations or other usage situations (see also Nielsen and Mourier 2007). On the *BLF* homepage, the user can choose

from six different user-driven situations, which correspond to six different types of extra-lexicographic needs (Figure 1).

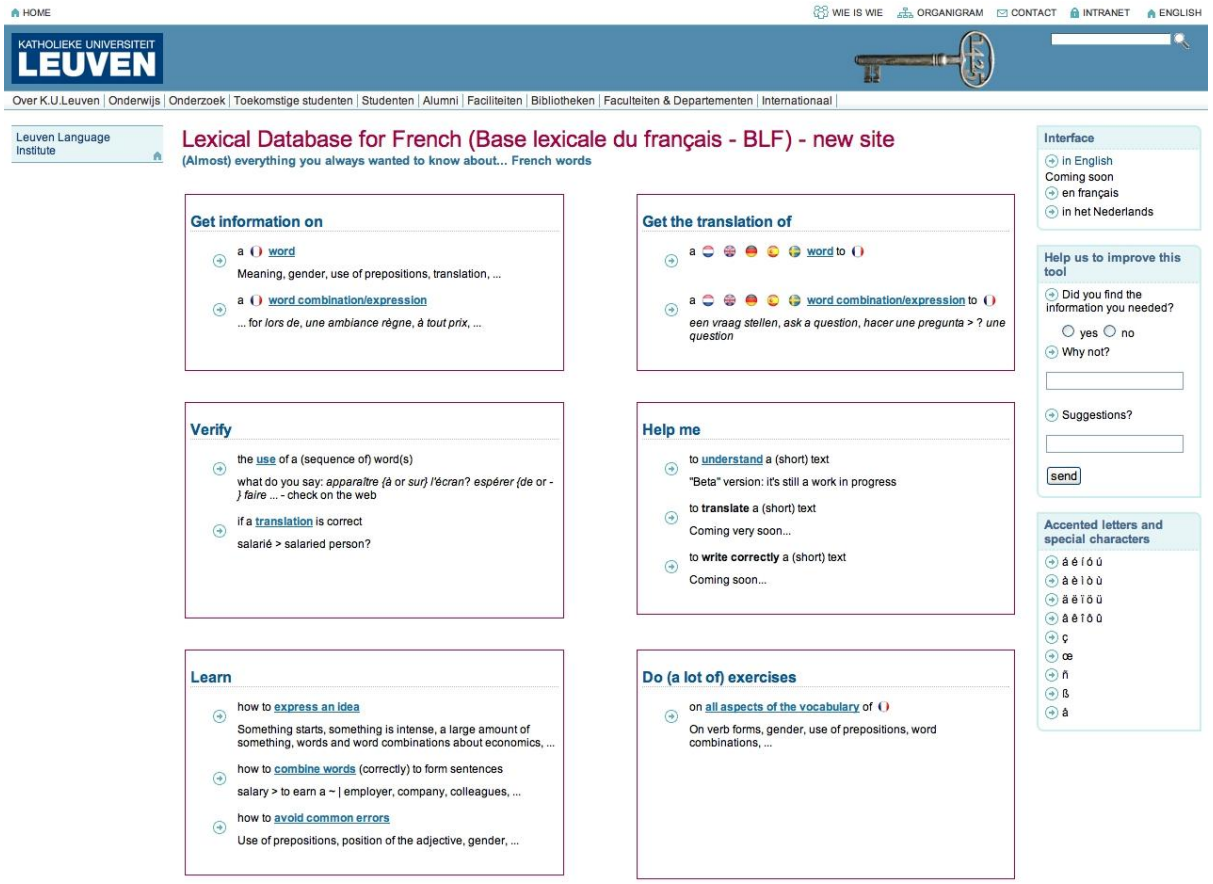


Figure 1.

1. Users who want information on a single word/multiword expression in the target language (box on the *BLF* homepage: 'Get information on ...').
2. Users who want the translation of a single word/multiword expression in their mother tongue to a target language ('Get the translation of ...').
3. Users who want to check the use of an expression/word combination or a translation ('Verify').
4. Users who want to learn/acquire the vocabulary of a foreign language (Learn).
5. Users who want to practice ('Do (a lot of) exercises').
6. Users who want the system to help them with general text reception, translation and production problems ('Help'). This part of the interface is still under construction (for more details, see Verlinde, Leroyer and Binon, forthcoming)

After the user has identified his specific consultation situation and submitted a word or word combination or selected a specific topic, they access a limited information item providing them with an answer to their specific question (Figure 2).

By adding an extended tracking and logging system to each form and link on all the *BLF* pages, any action taken by any user may be recorded and thus not only the words submitted can be tracked, but the complete look-up behaviour displayed by users as well; and a click-trail analysis may be performed, too.

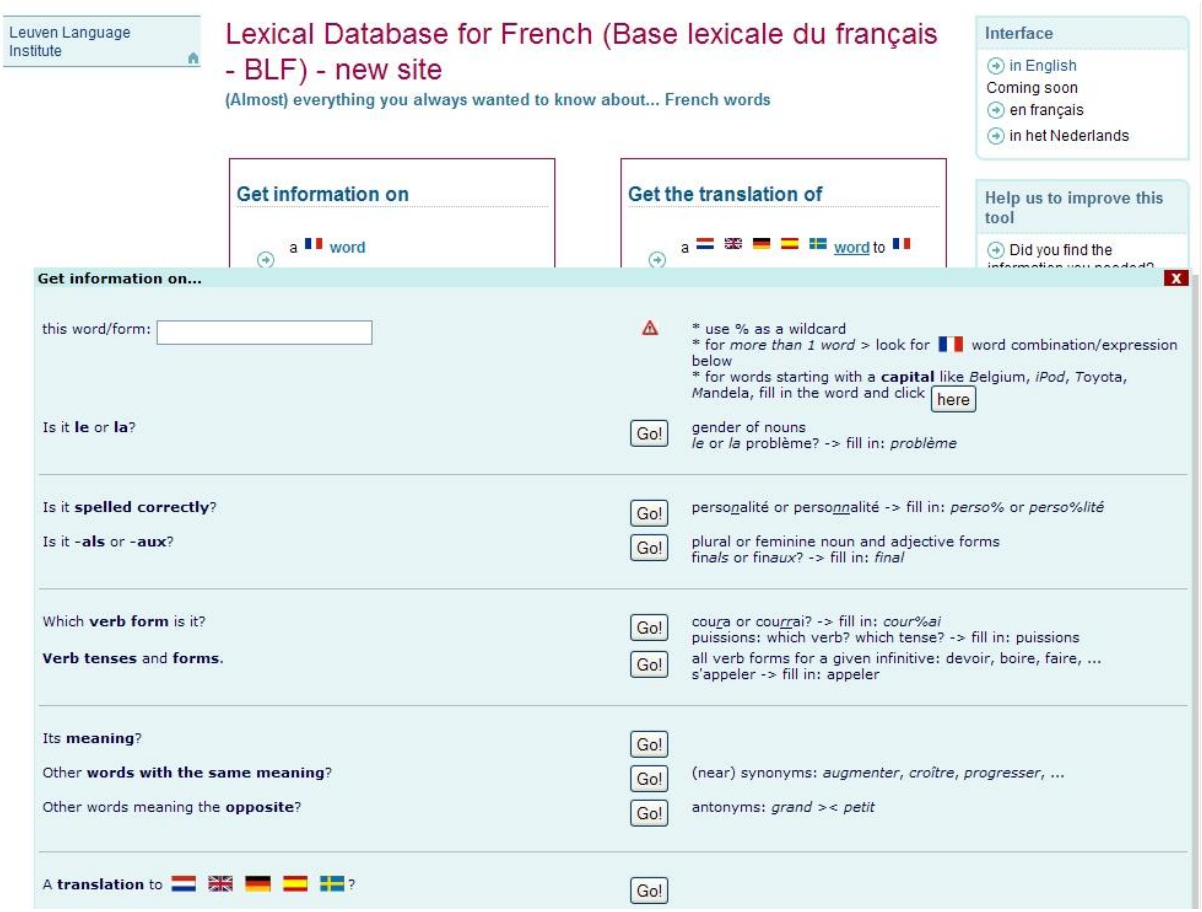


Figure 2.

The log files of the *BLF* contain information on

- submitted words and word combinations
- selections made on a webpage
- identification of the actual webpage
- date and time
- ip-address and name host server
- url of the previous and actual webpage
- session number

A session number is a unique number which is valid for the duration that the browser window is open.

3. On how the *BLF* is used

A closer look at the log files of the *BLF* reveals that 90.49% of all hits come from web spiders, Google webcrawlers for the most part. In this article, we will only consider 55752 hits from 'human' users.[1]

A first overview of results of the analysis of these 55752 hits refers to the way the *BLF* is browsed by the users.

[1] More common statistics on the number of look-ups/month, on the number of pages viewed by session, etc. and on the webcrawler queries may be found on http://ilt.kuleuven.be/blf/stats.

It is perhaps not surprising that almost 90% of all uses are related to the most regular dictionary uses: *get information on* (60.96%) and *get the translation of* (28.65%). The *learn* application is used in 7.35% of the consultations. This is a small, but encouraging percentage for a very specific tool the user is not familiar with from any other electronic dictionary.

The *get information on* consultation box allows the user to choose between about 20 information items on a word. Statistics show the prevalence of queries on meaning, gender, which is a typical problem for learners of French, and translation to another language (Graph 1).

Graph 1.

Maybe less expected are the high scores for syntax and lexical profile. Pronunciation, with sound files, is marginally selected, probably because the pronunciation problems for French are limited.

On the *BLF*, small information items are displayed, in answer to a specific question of a user. At any time, the user may click on links in order to get more general information on a word, mostly the whole article. Log files show, however, that users only want this extended information in no more than 11.23% of all cases. Dictionary consultations are thus notably targeted consultations, particularly for searches on translation (1.90%), gender (5.60%) and meaning (6.45%).

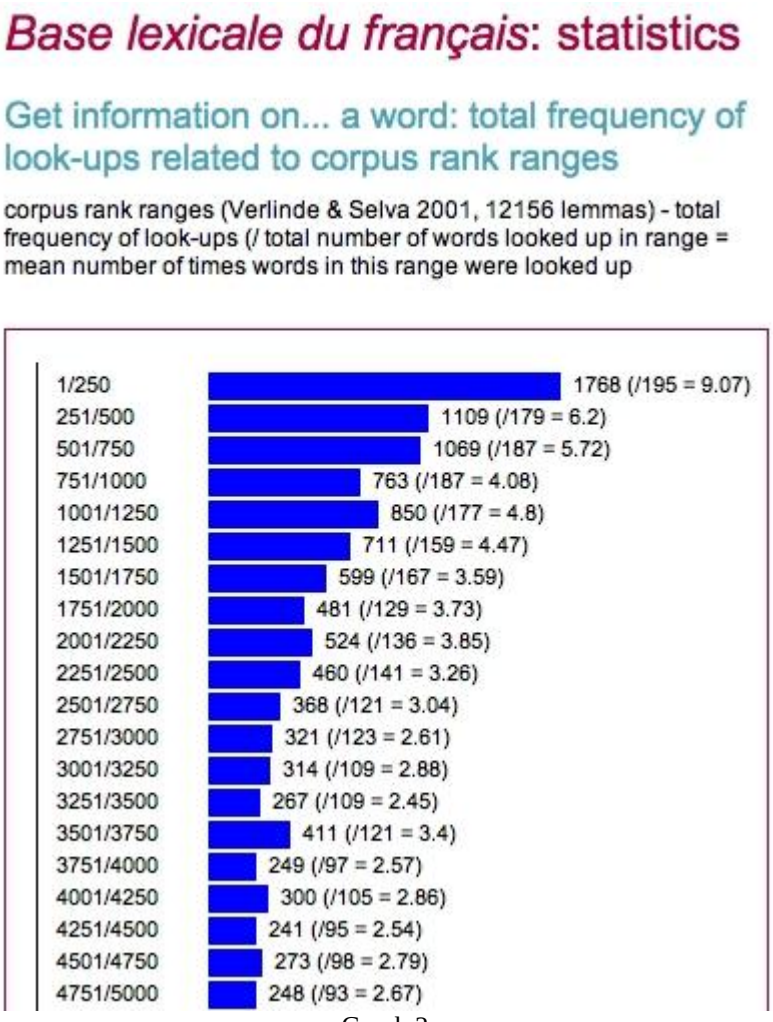
If we look at the most viewed pages, *get the translation of* a Dutch word to a French word is by far the most frequently looked up page, before the *get information on* a word's meaning or

gender. It is also noticeable that *learn* to combine words (lexical profile) is listed in the top 10 of most viewed pages.

Other statistics refer to the words submitted by the users.

In answer to the question if 'the top 100 searches [are] also the top 100 in a corpus' (De Schryver and Joffe (2004: 109), De Schryver (2006: 79) has previously shown that 'corpus frequencies do not predict look-up behaviour beyond the top few thousand words of a language'. Our data confirm these findings: the correlation between the corpus rank and the lemmatised look-up rank is very weak and does not exceed the value of 0.30, which is close to the values calculated by de Schryver (2006: 77). Our data do not provide 'support for the practice of including or omitting lemma signs in a dictionary based on frequency considerations (and by extension for corpus-based lexicography in general)'. However, the same analysis should be performed with data for each information item (meaning, synonyms, etc.) separately. Unfortunately, at this moment, we do not have enough data for each case to get reliable results.

The prevalence of the search on frequent words is clearly illustrated by the following graph (Graph 2).



Graph 2.

This graph shows that the most frequent words are most likely to be looked up (195 of the 250 words of the first range) and also most frequently looked up (a mean of 9,07 times for these 195 words).

If we look at the data in more detail, some user behaviour may be considered as quite surprising. We note for instance that more than 20% of the prepositions in the *BLF* have been looked up to control their syntax.

At this moment, with the limited amount of data, we are not able to perform more detailed analyses. We also think that supplementary filters on data will be necessary. A detailed analysis of the log files show for instance that some traces, when the same word is registered multiple) times at the same moment with similar host server identification, must be the consequence of a probable introductory session for students at a high school or university. We should also be able to filter out the cases where the user didn't start from the *BLF* homepage, for instance when he accessed the *BLF* directly from a Google search page result.

4. Towards an adaptive dictionary?

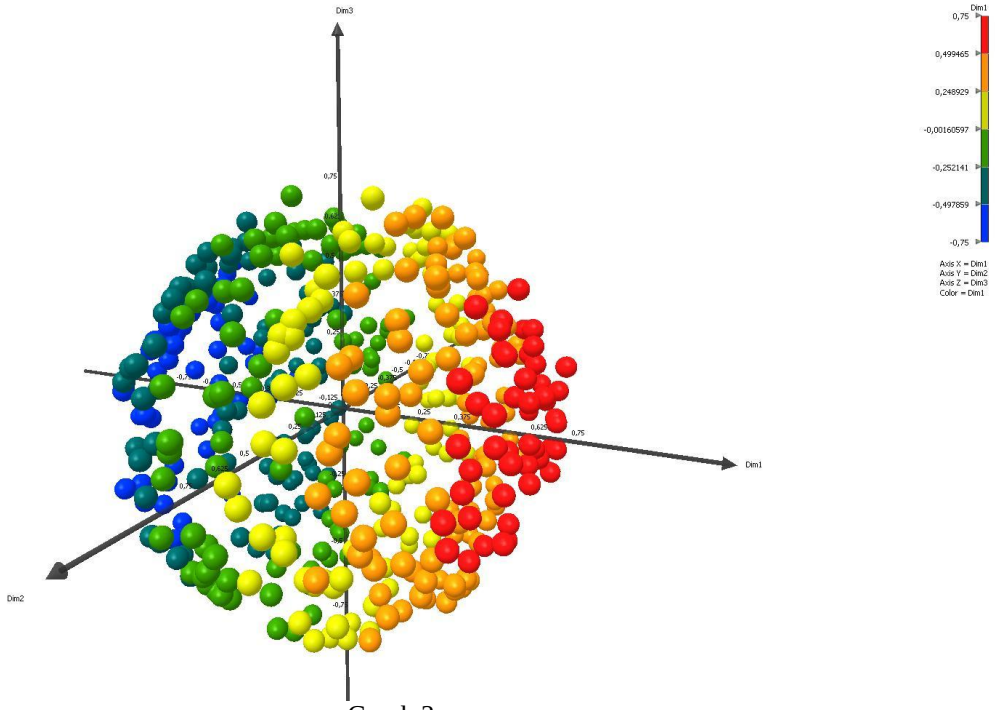
According to De Schryver (2004), numerous suggestions have been made to improve the user-friendliness of electronic dictionaries. The most basic idea of customisation is to display only part of the content of an article. But one could also think of a smart interface 'developing different profiles for different user situations' (Müller-Spitzer and Möhrs 2008: 44). However, developing such a smart interface is only possible if recurrent search strategies can be identified.

Therefore, we extracted from the log files the 2612 human accesses or sessions with at least three pages viewed. At least three accesses guarantee a more detailed consultation of the website and a minimal search strategy. We then selected 500 sessions at random, grouped in a 500-by-217 matrix. For each of the 500 users, a 0 (unaccessed) or 1 (accessed) value was given for each of the 217 different *BLF* web pages which were accessed at least once by the 500 selected users. We analysed this matrix using the multidimensional scaling statistic method, which is an exploratory statistical analysis assigning locations in a multidimensional space to observations – different dictionary users in this case – according to their similarity or dissimilarity, being the number of identical pages accessed by every user. [2]

The visual representation of the look-up behaviour of the 500 selected users in a three-dimensional space shows no pattern at all (Graph 3).

This implies that, as far as our data are concerned, in a natural setting, there are no frequent recurrent look-up strategies. Therefore, it will almost be impossible to conceive smart adaptive interfaces for dictionaries, unless more detailed data combining tracking data and other information as age or language level for instance, would eventually infirm this conclusion.

[2] The analysis has been carried out with XLSTAT statistical software.



Graph 3.

5. Perspective

In a dictionary almost anyone may look-up almost anything almost anywhere. This could be the overall conclusion of the study we carried out on the detailed log files of the *BLF*. We believe attempts to customize the interface of a dictionary to look-up behaviour (dreams of 'smart' customization in De Schryver 2003: 185-186) cannot be very successful, as illustrated by the diversity of dictionary usages. Maybe we have to revisit the dictionary and provide a tool which is able to deal with the fact 'that users increasingly assume that electronic dictionaries behave like Web search engines such as *Google'*, and allow them to 'type in concatenations of keywords, combinations and phrases surrounded by quotes, entire sentences, and even dump full paragraphs (lifted from other sources) into the search field' (De Schryver et al. 2006: 71). This has been attempted in the *Help* application in the *BLF*, to which NLP-inspired analysis tools could be added. On the other hand, according to Chon (2009: 49-50), lexicographers should think about ways to encourage users to 'cross-reference their findings when the information retrieved is incomplete or doubtful, which can easily be facilitated by interfacing'. This verification strategy should become inherent to any search dealing with word combinations for instance. At this moment, on the *BLF* website, only 1.43% of all viewed page have to do with verifying.

6. Conclusion

The analysis of the log files of the *BLF* suggests a diversity of dictionary look-up strategies. Different ways of customization of the dictionary interface may thus not give the expected results. We may conclude that the challenge for the lexicographer is to find how actual (NLP) technology could be combined with all possible resources available on the internet (monolingual and bilingual corpora) to create a fully operational 'integrated lexical information system' (Heid 2008: 147).

References

Awstats. http://awstats.sourceforge.net/ [access date: 17 Feb. 2010]

Base lexicale du français. http://ilt.kuleuven.be/blf [access date: 17 Feb. 2010]

Bogaards, P. (2003). 'Uses and users of dictionaries'. In van Sterkenburg, P. (ed.). *A practical guide to lexicography*. Amsterdam/Philadelphia: John Benjamins. 26-33.

Cambridge Advanced Learner's Dictionary online, top 40 words. [Cambridge University Press] http://dictionary.cambridge.org/top40/top40.asp [access date: 17 Feb. 2010]

Chon, Y.V. (2009). 'The electronic dictionary for writing: a solution or a problem?' In *International Journal of Lexicography* 22 (1). 23-54.

De Schryver, G.-M. (2003). 'Lexicographers' dreams in the electronic-dictionary age'. In *International Journal of lexicography* 16 (2). 143-199.

De Schryver, G.-M.; Joffe, D. (2004). 'On how electronic dictionaries are really used'. In Williams, G.; Vessier, S. (eds.). *Proceedings of the Eleventh EURALEX International Congress*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 187–196.

De Schryver, G.-M.; Joffe, D.; Joffe, P.; Hillewaert, S. (2006). 'Do dictionary users really look up frequent words ? – On the overestimation of the value of corpus-based lexicography'. In *Lexikos* 16. 67-83.

Google Analytics. [Google Inc.] http://www.google.com/analytics/ [access date: 17 Feb. 2010]

Heid, U. (2008). 'Corpus linguistics and lexicography'. In Lüdeling, A.; Kytö, M. *Corpus linguistics. An International Handbook*. Berlin/New York: W. de Gruyter. 131-153.

Logaholic. [Logaholic Web Analytics] http://www.logaholic.com/wp/ [access date: 17 Feb. 2010]

Měchura, M. B. (2008). 'Giving them what they want: search strategies for electronic dictionaries'. Bernal, E.; DeCesaris, J. (eds.). *Proceedings of the XIII EURALEX Internal Congress*. Barcelona: Institut Universitari de Lingüistica aplicada, Universitat Pompeu Fabra.

Müller-Spitzer, C.; Möhrs, C. (2008). 'First ideas of user-adapted views of lexicographic data exemplified on OWID and elexiko'. In Zock, M.; Huang, C.-R. (eds.). *Coling 2008: Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*. Manchester.

Nielsen, S.; Mourier, L. (2007). 'Design of a function-based internet accounting dictionary'. In: Gottlieb, H.; Mogensen, J.E. (eds.). *Dictionary visions, research and practice. Selected papers from the 12th international symposium on lexicography, Copenhagen 2004*. Amsterdam/Philadelphia: John Benjamins. 119-135.

TshwaneLex. [TshwaneDJe HLT] tshwanedje.com/tshwanelex/ [access date: 17 Feb. 2010]

Verlinde, S.; Selva, T. (2001). 'Nomenclature de dictionnaire et analyse de corpus'. In *Cahiers de lexicologie* 79. 113-139.

Verlinde, S.; Leroyer, P.; Binon, J. (2010). 'Search and you will find. From stand-alone lexicographic tools to user driven task and problem-oriented multifunctional leximats'. In *International Journal of Lexicography* 23 (1).

Welker, H. A. (2006). *O Uso de dicionários: Panorama geral das pesquisas empíricas*. Brasilia: Thesaurus.

XLSTAT. [Addinsoft] www.xlstat.com/ [access date: 17 Feb. 2010]