Dictionary Users do Look up Frequent and Socially Relevant Words. Two Log File Analyses.

Sascha Wolfer, Alexander Koplenig, Peter Meyer, Carolin Müller-Spitzer
Institut für Deutsche Sprache, Mannheim, Germany
wolfer@ids-mannheim.de

Abstract

We start by trying to answer a question that has already been asked by de Schryver et al. (2006): Do dictionary users (frequently) look up words that are frequent in a corpus. Contrary to their results, our results that are based on the analysis of log files from two different online dictionaries indicate that users indeed look up frequent words frequently. When combining frequency information from the Mannheim German Reference Corpus and information about the number of visits in the Digital Dictionary of the German Language as well as the German language edition of Wiktionary, a clear connection between corpus and look-up frequencies can be observed. In a follow-up study, we show that another important factor for the look-up frequency of a word is its temporal social relevance. To make this effect visible, we propose a de-trending method where we control both frequency effects and overall look-up trends.

Keywords: research into dictionary use; frequency; corpus; social relevance; log file analysis

1 Introduction[1]

In this paper, we use the 2012 log-files of two German online dictionaries (Digital Dictionary of the German Language and the German language edition of Wiktionary) and the 100,000 most frequent words in the Mannheim German Reference Corpus (Deutsches Referenzkorpus, DEREKO) from 2009 (Kupietz et al., 2010) to answer the question of whether dictionary users really do look up frequent words, first asked by de Schryver et al. (2006). The research question standing behind is whether it actually makes sense to select words based on frequency, or, in other words, if it is a reasonable strategy to prefer words that are more frequent over words that are not so frequent. Answering this question is especially important when it comes to building up a completely new general dictionary from scratch and the lexicographer has to compile a headword list. By using an approach to the comparison of log-files and corpus data which is completely different from that of the aforementioned authors, we provide empirical evidence that indicates – contrary to the results of de Schryver et al. and Verlinde & Binon (2010) – that the corpus frequency of a word can indeed be an important factor in determining

1 In Koplenig, Meyer, & Müller-Spitzer (2014) we present and discuss the results of this study in more detail.

what online dictionary users look up. In addition, we incorporate word class information readily available in Wiktionary into our analysis to improve our results considerably. In a follow-up study, we show that (temporal) social relevance of particular words can influence look-up behaviour considerably. For the latter study, we used the 2013 log files of the German language edition of Wiktionary.

2 Previous research

To understand whether including words based on frequency of usage considerations makes sense, it is a reasonable strategy to check whether dictionary users actually look up frequent words. Of course, in this specific case, it is not possible to design a survey (or an experiment) and ask potential users whether they prefer to look up frequent words or something like that. That is why de Schryver and his colleagues (2006) compared a corpus frequency list with a frequency list obtained from log-files. The aim of De Schryver et al.'s study was to find out whether dictionary users look up frequent words. Due to the nature of the statistical method they used, de Schryver et al. (2006) actually tried to answer two different questions: do dictionary users look up frequent words frequently? And, do dictionary users look up less frequent words less frequently? (cf. Koplenig, Meyer, & Müller-Spitzer, 2014: 232). The result of their study is part of the title of their paper: "On the Overestimation of the Value of Corpus-based Lexicography". Verlinde & Binon (2010: 1148) replicated the study of de Schryver et al. (2006) using the same methodological approach and essentially came to the same conclusion.[2]

In this paper, we will try to show why de Schryver et al.'s straightforward approach is rather problematic due to the distribution of the linguistic data that is used. In this context we suggest a completely different approach and show that dictionary users do indeed look up frequent words (sometimes even frequently). In a follow-up study, we present a case study that suggests that, as soon as frequency information is partialled out, analyses of log-files can also reveal information about the (sometimes very short-lived) social importance of particular words.

3 The Data

Corpus data

When we look at the DEREWO list, a word list compiled using the DEREKO corpus, and plot the relative frequency against the rank, we receive a typical Zipfian pattern. This means that we have a handful of word forms that have a very high frequency and an overwhelming majority of word forms that

2 In contrast, **Henrik** Lorentzen, Nicolai H. Sørensen and Lars Trap-Jensen in their talk at the e-lexicography conference 2013 also came to the conclusion that frequent words in a corpus are also frequently looked up in a dictionary (Talk: "An odd couple – corpus frequency and look-up frequency: what relationship?" Video available at http://eki.ee/elex2013/videos/ [last access on 02/04/2014].

have a very low frequency. Or, in other words, our DEREWO list consists of 3,227,479,836 word form tokens. The 200 most frequent word form types in the list make exactly half of those tokens.

Log-files

The Wiktionary log file types are roughly 8 w times as many as the DWDS log file types. To make the results both comparable and more intuitive, we rescaled the data by multiplying the raw frequency of a query by 1,000,000, dividing it by the sum of all query tokens and rounding the resulting value. We then removed all queries with a value smaller than one. Thus, the resulting variable is measured in a unit that we would like to call *poms* (per one million searches). For example, a value of 8 means that the corresponding phrase is searched for 8 times per one million search requests. Table 1 summarizes the resulting distribution.

Category (*poms*)	**Wiktionary log-files (%)**	**DWDS log-files (%)**
1	57.94	57.30
2 - 10	33.71	31.15
11 - 49	6.69	9.09
50 - 500	1.63	2.44
500 +	0.03	0.02
Total	**100.00 (abs. 185,071)**	**100.00 (abs. 156,478)**

Table 1: Categorized relative frequency of the log file data.

Category	**X searches *poms***	**Wiktionary log-files (%)**	**DWDS log-files (%)**
regular	at least 1	100.00	100.00
frequent	at least 2	42.06	42.70
very frequent	at least 11	8.35	11.55

Table 2: Definition of the categories used in the subsequent analysis and relative log file distribution.

Table 1 shows two things: firstly, the Wiktionary and the DWDS log-files are quite comparable on the poms-scale; secondly, just like the corpus data, the log-files are heavily right skewed. More than half of all query types consist of phrases only searched for once poms. When we cumulate the first two categories, we can state for both the Wiktionary and the DWDS data that 90% of the queries are requested 1 up to 10 times poms. So there is only a small fraction of all phrases in the log-files that are searched for more frequently.

4 Data analysis

In the previous section, we described the data and presented a new unit of measurement called *poms*. If we think about our research question again – whether dictionary users look up frequent words (frequently) – it is necessary to find an appropriate method for analyzing the data using this unit. For example, we could regress the log file frequency (in poms) on the corpus frequency, but an ordinary least squares (OLS) regression implies a linear relationship between the explanatory and the response variable, which is clearly not given. (Log-)Transforming both variables does not solve our problem, either, and this is in any case seldom a good strategy (O'Hara & Kotze, 2010). We could use the appropriate models for count data such as Poisson regression or negative binomial regression, but, as Baayen (2001, 2008: 222-236) demonstrates at length, we still have to face the problem of a very large number of rare events (LNRE), which is typical for word frequency distributions. And even if we could fit such a model, it would remain far from clear what this would imply for our initial lexicographical question. Using the standard Pearson formula to correlate the corpus and the log-file data suffers from the same nonlinearity problem as the OLS approach. Therefore de Schryver et al. (2006) implicitly used the nonparametric Spearman rank correlation coefficient which is essentially just the Pearson correlation between ranked variables. We believe that this is still not the best solution, mainly because, on a conceptual level, ranking the corpus and log-file data implies that subsequent ranks are equidistant in frequency, which is clearly not the case. Again, the inherent Zipfian character of the distribution explains why the ranks are far from equidistant. For example, the difference in frequency between the first and the second rank is 251,480, whereas the difference between the 3000th and 3001th is only 5. Nevertheless the Spearman rank correlation coefficient treats the differences as equal[3].

In the last section, we grouped the log-files (cf. Table 1) into poms categories. As a possible solutions to the problems we just outlined, we now use this grouping again and stipulate the following categories: if a word form is searched for at least once poms, it is searched for regularly, if it is searched for at least twice, we call it frequent, and if it is searched for more than 10 times, it is very frequently searched for. Table 2 sums up the resulting values. Please keep in mind that according to this definition, a very frequent search term also belongs to the regular and the frequent categories. Our definition is, of course, rather arbitrary and mainly has an illustrative function, but due to the Zipfian distribution of the data, only a minority of the searches (roughly 4 out of 10) occur more than once poms and even fewer words (roughly 1 out of 10, roughly 8 percent for Wiktionary, roughly 12 percent for the DWDS) are searched for more than ten times poms (cf. Table 1). Therefore this definition at least approximates the distribution of the log file data. Nevertheless, instead of using the categories presented in the first column of Table 2, we could also use the second column to label the categories. So it must be borne in mind that the labels merely have an illustrative function.

3 In principle, we could use another similarity metric, for example the cosine measure (i.e. the normalized dot product, cf. Jurafsky & Martin, 2009: 699), but as in the case of using a count regression model, we are not sure what the value of the coefficient would actually imply both theoretically and practically.

We then wrote a Stata program that starts with the first ten DEREKO ranks and then increases the included ranks one rank at a time. At every step, the program calculates how many of the included word forms appear in the DWDS and Wiktionary log-files regularly, frequently, and very frequently (scaled to percentage). Table 3 summarizes the results for 6 data points.

Included DEREKO ranks	DWDS (%)			Wiktionary (%)		
	regular	frequent	very frequent	regular	frequent	very frequent
10	100.0	100.0	100.0	100.0	100.0	100.0
200	100.0	99.0	87.5	99.5	99.5	86.5
2,000	96.9	91.0	67.6	98.4	96.0	64.9
10,000	85.5	72.9	47.5	86.3	75.3	40.2
15,000	80.3	66.5	41.8	77.4	66.1	33.7
30,000	69.4	54.6	31.3	62.7	50.9	23.4

Table 3: Relationship between corpus rank and log file data.

In this table, the relationship between the corpus rank and the log file data becomes obvious: the more DEREKO ranks we include, the smaller the percentage of those word forms appearing regularly/frequently/very frequently in both the DWDS and the Wiktionary log-files. Let us assume for example that we prepare a dictionary of the 2,000 most frequent DEREKO word forms; our analysis of the DWDS and the Wiktionary data tells us that 96.9 % of those word forms are searched for regularly in DWDS, 91.0 % are searched for frequently and 67.6 % are searched for very frequently. For Wiktionary, these figures are a bit higher.

Figure 2 plots this result for the DWDS and the Wiktionary log-files separately. It comes as no surprise that the curve is different for the three categories, being steepest for the very frequent category, since this type of log file data only makes up a small fraction of the data. To further improve our analysis, we looked at the word forms that are absent in both the DWDS and the Wiktionary log-files but that are present in the unlemmatised DEREKO corpus data. There is a roughly 60% overlap, which means that 6 out of ten word forms missing in the DWDS data are also missing in the Wiktionary data. To understand this remarkable figure, we tried to find out more about the words that are missing in the log-files but are present in the corpus data. In our talk, we can present how we used this data to improve our analyses.

We would like to provide an additional impression of our results by asking what proportion of all search requests (tokens) could be covered with such a corpus-based strategy. If we again use the example of the first 15,000 DEREKO most frequent word forms, then around half of all DWDS search requests that occur regularly or frequently (poms) are covered, while around two-thirds of all very frequent requests are successful. If we included the 30,000 most frequent DEREKO words, roughly two-thirds of the regular and frequent and 80.0% of the very frequent DWDS search requests would be covered in

the dictionary. In other words, this means if we included the 30,000 most frequent DEREKO word forms, the vast majority of requests would be successful.

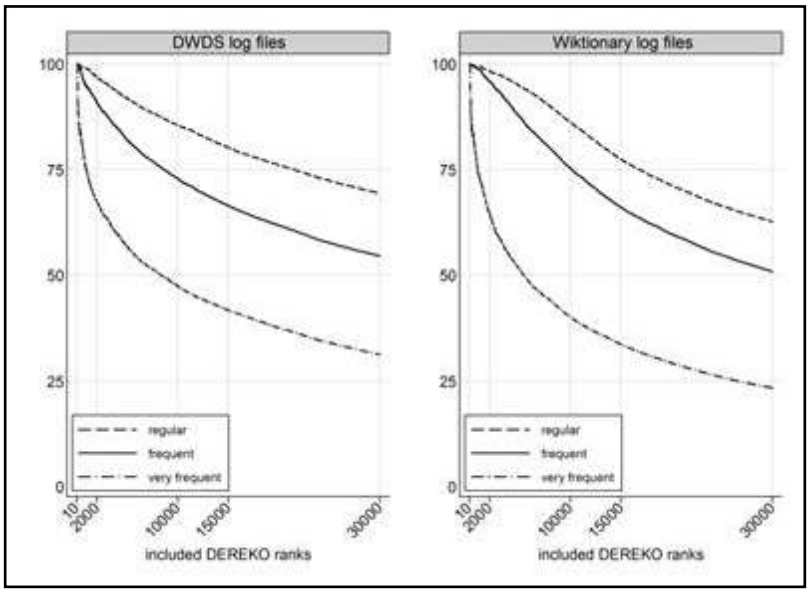

Figure 1: Percentage of search requests which appear in the DWDS/Wiktionary log-files as a function of the DEREKO rank.

5 Social relevance and dictionary usage

The previous sections showed why we think that it is a good idea to include frequently occurring words in dictionaries. Although it is clear that there is a strong and reliable relationship between frequency of occurrence and look-up frequency, it is also quite clear that the former is not the *only* predictor of the latter. To investigate further contributing factors, we consulted log files from the German Wiktionary in the year 2013 and aggregated the hourly Wiktionary log files to weekly datasets to keep the computations manageable. We excluded all pages with titles longer than 80 characters. We also excluded all pages that were visited fewer than one time in one million visits. The overall aim of this study is to identify points in time where certain words are looked up extraordinarily often. To achieve this, we need to control for the overall trend of look-up frequency of each word. It is no surprise that look-up frequency varies over time. Words are looked up more or less often during the course of a year. This variation can be captured by the overall trend *within* a word's visits. By controlling for these long-term trends, we also capture general look-up differences *between* words that stem i.a. from the frequency effects outlined above. What we are currently interested in are rather short-term 'peaks' in the number of visits a specific word receives. The number of visits a specific word receives is the sum of the overall trend for this word and 'noise' which is not captured by this trend (cf. Becketti, 2013:92-95,103,109). This noise, or – informally speaking – what is left over after the overall trend has been considered is exactly the kind of data we are interested in. To extract this variable, we fitted a Tukey

smoother using running medians of length 3[4]. This smoother captures the trend. The variable we are going to use is the difference between this smoother and the actual visits, we call this the difference score or residual visits. Using this technique, we can look beyond the effect of frequency and overall tendencies of a specific word. In other words, this technique enables us to identify extraordinary look-up behavior for individual words in individual points in time[5]. To extract interesting words, we rank words by their smooth-difference score. All highly ranking words have especially high proportions of unexplained variance in visits per one million visits in the respective week.

We will now describe two headwords with noticeable difference scores to provide a first impression of our results. The word "Furor" (English "furor", "rage") takes rank 14 of the ordered list of difference scores. The differences between smoothed and observed visits per one million visits are constantly around zero. However, in week 10 of 2013 (04/03/2013 to 10/03/2013), its difference scores go up to 2,210 with a total of 4,687 visits (for all other weeks except week 10, the mean of raw visits for "Furor" is 60.7). In this week, German president Joachim Gauck used the word "Tugendfuror" (roughly: "furor/rage of virtue") in a debate on sexism in Germany. His whole comment, and especially the word "Tugendfuror" was subject of public discussion in Germany throughout the media. Figure 3 shows the difference scores of "Furor" on a daily basis for one month (20/02/2013 to 20/03/2013). One can clearly see how residual searches *poms* rise for one critical day (06/03/2013) and then take one to two days to 'normalize' again to a residual value around zero.

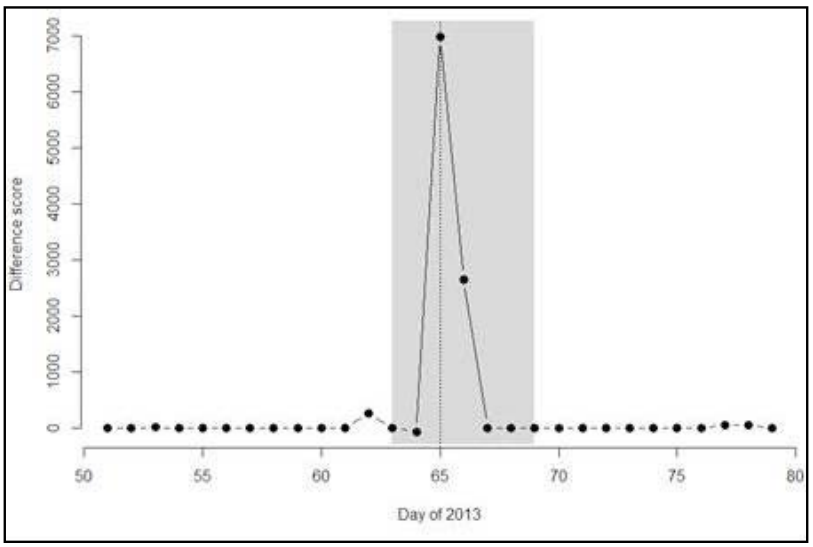

Figure 2: Difference between smoothed and observed visits per 1 million visitsfor "Furor" between 20/02/2013 and 20/03/2013. Week 10 is highlighted.

4 To do this, we used the default behavior of the function *smooth()* provided by the 'stats' package of the statistical programming language R (R Core Team, 2013).

5 Of course, these differences can also take negative values. Indeed, many of them do. This means that a word received less searches poms in a particular week than would be expected given the word's overall trend. If we normalize these differences by dividing the difference score by the smoothed visits per one million visits, we can also compare words between one another. In this paper, we did not apply this operation.

Interestingly, Joachim Gauck used the word in an interview[6] already published on 03/03/2014. Though one can see a minimal rise on that day, it took three days until other major newspapers picked up the debate because other voices alluded to potential problems of Gauck's choice of words[7]. Obviously, quite a lot people then wondered what the head of the compound "Tugendfuror" actually means and referred to Wiktionary during those days. "Furor" is a good example of a temporarily socially relevant word. Actually, subject of discussion was the lexical meaning of "Furor" and its connections to other, potentially pejorative words. So, the discussion was lexico-semantic in nature and it comes as no surprise that people tended to look up the word the German president used.

However, there are other noticeable words in certain periods of time, which are not directly related to discussions in society or politics that are lexical in nature. Figure 4 shows the residual searches poms of the word "Borussia" in time. "Borussia" is latin for "Prussia" and part of the name of several German sports clubs. The most prominent ones are football clubs.

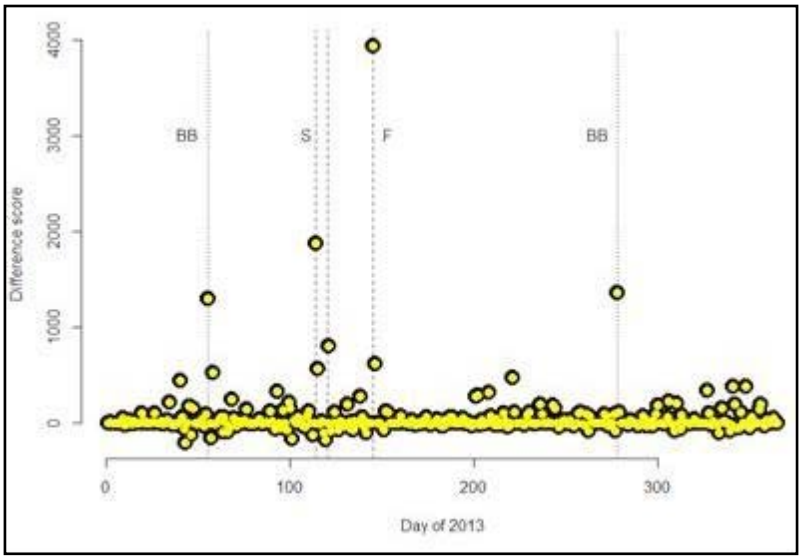

Figure 3: Difference between smoothed and observed visits per 1 million visits for "Borussia" for the whole year 2013. Vertical lines indicate football matches (BB: Borussia Mönchengladbach vs. Borussia Dortmund, S: UEFA Champions League semi-finals, F: Final).

Peaks are identifiable in the difference scores for "Borussia" over time; symbolizing temporarily increased look-ups for "Borussia" in Wiktionary that cannot be explained by frequency of occurrence or overall search preferences alone. Each dashed vertical line in Figure 4 represents one match in the knockout phase of the UEFA Champions League (CL) competition with the participation of Borussia Dortmund. Look-ups of "Borussia" sharply increase around match days. For the semi-finals ("S") and especially the all-important final match ("F"), residual searches poms increase sharply around match

6 See http://www.spiegel.de/politik/deutschland/sexismus-debatte-gauck-beklagt-tugendfuror-im-fall-brueder-le-a-886578.html [last access on 01/04/2014].

7 See http://www.sueddeutsche.de/politik/sexismus-debatte-als-tugendfuror-aufschrei-wegen-gauck-1.1616310 [last access on 02/04/2014].

days. There are two other vertical lines ("BB") which do not mark a match day in the CL. BB marks 24/02/2013 and 05/10/2013, the days Borussia Dortmund competed against Borussia Mönchengladbach in the German first division. This match is associated with increased difference scores, too. In contrast, no other match in the German first division did lead to increased residual searches poms for "Borussia". Obviously, the popularity and importance of the CL competition led to repeated increases in the social relevance of the term "Borussia". Also, when both Borussias competed against each other in the national championship, public interest in the somewhat cryptic name part was also increased. In comparison to the "Furor" case presented above, the look-up behavior concerning "Borussia" is more surprising. There is no lexico-semantic debate involved that could persuade people to look up "Borussia". Increased media coverage and general public awareness concerning a football club alone seems sufficient to trigger noticeable increases in look-up behavior. This is a remarkable and important observation for research into dictionary use. Another example is the word "larmoyant" (English "lachrymose" meaning "tearfully sentimental") which was used in a sports commentary in an exhibition match between the French and German football national team. Here, the commentator described one specific German national as being too "larmoyant" which led to sharply increased lookups within the same hour (which is the minimum temporal resolution available for the Wiktionary log files).
There are several more interesting cases extractable from the Wiktionary log files that we cannot report here. Social relevance in other cases was induced by a variety of social contexts like TV game shows and even astronomical events like a solstice. Explaining why residual look-ups increased in a specific timeframe is interesting and it definitely points to the fact that the social context directly influences look-up frequencies in internet dictionaries – all in a very short time frame. In future research, however, we want to operationalize social relevance of words in a large-scale, automatized way. Such a measure would enable us to correlate these two measures not only over singular cases but many words. Furthermore, one could identify social contexts that are especially capable (or others that are not capable at all) to trigger look-up peaks in online dictionaries. This line of research could also contribute to the overall question of this paper: Which words should be included in dictionaries? Certainly, words that are socially highly relevant over long periods of time are good candidates.

6 Conclusion

In general, the use of a corpus for linguistic purposes is based on one assumption:
"It is common practice of corpus linguistics to assume that the frequency distributions of tokens and types of linguistic phenomena in corpora have – to put it as generally as possible – some kind of significance. Essentially more frequently occurring structures are believed to hold a more prominent place, not only in actual discourse but also in the linguistic system, than those occurring less often" (Schmid, 2010: 101).

We hope that we have provided evidence which shows that, based on this assumption, corpus information can also be used fruitfully when it comes to deciding which words to include in a dictionary. This corpus-based strategy is no "magic answer". We simply think it is the best one there is, given that there are no other systematic alternatives.

Beyond that, social relevance of words or their (extraordinary) presence in social discourse seems to be a highly relevant factor in this context.

7 References

Baayen, R. H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R.* Cambridge, UK: Cambridge University Press.

Becketti, S. (2013). *Introduction to Time Series Using Stata*. College Station: Stata Press.

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. *Lexikos, 16*, 67–83.

Jurafsky, D. & Marti, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational Linguistics, and speech recognition*. Upper Saddle River: Pearson Education (US).

Kupietz, M., Belica, C., Keibel, H., & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari, D. Tapias, M. Rosner, S. Piperidis, J. Odjik, J. Mariani, ... K. Choukri (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-10)* (pp. 1848–1854). Valetta, Malta: European Language Resources Association (ELRA).

Koplenig, A., Mayer, P., & Müller-Spitzer, C. (2014). Dictionary users do look up frequent words. A log file analysis. In: C. Müller-Spitzer (Ed.). *Using online dictionaries* (pp. 229-250). Berlin, New York: de Gruyter. (Lexicografica: Series Maior 145).

O'Hara, R.B. & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution, 1*(2), 118-122.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org (last accessed 02/04/2014).

Schmid, H.-J. (2010). Does frequency in text instantiate entrenchment in the cognitive system? In D. Glynn & K. Fischer (Eds.), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches* (pp. 101–133). Berlin, New York: de Gruyter.

Verlinde, S., & Binon, J. (2010). Monitoring Dictionary Use in the Electronic Age. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV Euralex International Congress* (pp. 1144–1151). Ljouwert: Afûk.

Acknowledgements

We are very grateful to the DWDS team for providing us with their log-files.