Computational Linguistics in the Netherlands Journal 4 (2014) 191-208 Submitted 06/2014; Published 12/2014

# BasiLex: an 11.5 million words corpus of Dutch texts written for children

**Agnes Tellings*** A.Tellings@ru.nl
**Micha Hulsbosch*** m.hulsbosch@pwo.ru.nl
**Anne Vermeer**** anne.vermeer@uvt.nl
**Antal van den Bosch*** a.vandenbosch@let.ru.nl

* *Radboud University, Nijmegen, The Netherlands*

** *Tilburg University, Tilburg, The Netherlands*

## Abstract

This article discusses Basilex, a 13.5-million tokens, 11.5-million Dutch words corpus of written language offered to children in the elementary school age, which was recently finalized. The corpus is automatically analyzed at the levels of part-of-speech tagging and lemmatization, and a limited amount of polysemous words has been partly automatically disambiguated. Also, a lemma-based lexicon is derived. The aim of the present article is threefold: First, to give a description of BasiLex and how it was built, and to discuss its validity. Second, to compare the BasiLex lexicon with two other lexicons regarding differences in their most frequent words: the Schrooten and Vermeer (1994) lexicon, a small and now outdated Dutch corpus of language addressed to children, and a derived lexicon of SoNaR, an adult written language corpus (Oostdijk et al. 2013). Third, we discuss some potential educational applications of BasiLex.

## 1. Introduction

Large annotated language corpora and associated lexicons and frequency lists are of great and increasing interest to researchers in the field of cognitive linguistics, computational linguistics, sociolinguistics, psycholinguistics, language development, and related areas. For the Dutch language, the oldest word frequency list dates from the mid-1930s (De la Court 1937). The list was manually gathered on cards, containing word counts in a corpus of one million words of which 25% consisted of children's books. In the second half of the 20th century, computers started to be used, as in Van Berckel (1962), Martin (1967) and Uit den Boogaart (1975). From the early 1990s onward increasingly large-scale digital lexicons have been produced, along with increasingly comprehensive corpora, such as CELEX (Baayen et al. 1993); see http://tst-centrale.org/nl/producten for a non-exhaustive list.

Most of these corpora comprise spoken and written language by and directed at adults. An exception is the Jasmin-CGN corpus (Cucchiarini et al. 2006), which addresses the need to have spoken language corpora produced by children, non-native speakers of Dutch, and elderly people, in the Netherlands and Flanders. Yet, there are still no large corpora of written language that children receive and produce, especially for educational purposes or for research regarding children's spoken and written language development and their developing mental lexicon. Until recently, only two such corpora existed: a child output corpus of 450,000 tokens with an associated list of 4,332 lemmas of ethnic minority children in elementary school (Coenen and Vermeer 1988), and a child input corpus of 1.9 million tokens and a list of 24,844 lemmas (Schrooten and Vermeer 1994). From the latter corpus of child-directed speech (540,000 tokens) and written language offered to children in Dutch elementary school (children's literature, school texts) a lexicon of 15,000 words with frequency measures was derived. Different target lists (i.e., lists of words children should know) have been developed based on the Schrooten and Vermeer (1994) corpus and lexicon and

on limited further research, some of these lists partly being based on each other (for an overview see Bacchini et al. 2005). As the basis for these lists is quite small and outdated, there was an urgent need for a larger and more recent corpus and derived lexicon of language written for children. One of the other reasons for building a large and up-to-date child input corpus and lexicon is the quite recent focus in (psycho)linguistic and computational linguistic research on multi-word constructions as being important building blocks of the mental lexicon in addition to single words (see e.g. Wray 2002, Bannard and Matthews 2008, Elman 2009, Arnon and Snider 2010, Frank et al. 2012). The study of these multi-word constructions requires corpora next to single-word lexicons and target lists.

To fill these needs, in May 2014 the BasiLex corpus was finalized, subsidized by the Dutch government: a 13.5-million tokens, 11.5-million words corpus of written language offered to children in the elementary school age. The corpus is automatically analyzed at the levels of part-of-speech tagging and lemmatization, and 5184 polysemous words have been partly automatically and partly manually disambiguated. Also, a lemma-based lexicon is derived. The aim of the present article is threefold. First, to give a description of BasiLex and how it was built and to discuss its validity. We do this in section 2. Second, to compare the BasiLex lexicon with the Schrooten and Vermeer (1994) lexicon and with an adult written language lexicon, focusing on the most salient differences in words frequencies. For the latter we use frequency lists derived from SoNaR, a large adult Dutch written language corpus (Oostdijk et al. 2013). We do this in Sections 3 and 4. The comparisons with Schrooten & Vermeer on the one hand and with SoNaR on the other hand are valuable for various reasons. First, if BasiLex is indeed a reasonable representation of contemporary written language input of elementary school children, there should be considerable frequency differences between words that were current some twenty years ago (and which should have high relative frequencies in Schrooten & Vermeer) but nowadays less so. Second, if children's and adults' written language input differ, there should be considerable frequency differences in BasiLex versus SoNaR for words we would typically expect to be adult language, and similarly for words we would typically view as child language. However, since children and adults, obviously, live in the same language world for the larger part, we also expect overlap between word frequencies in BasiLex and SoNaR. In the final part of our paper, Section 5, we discuss potential educational applications of BasiLex.

## 2. The BasiLex corpus and derived lexicon

For BasiLex we collected materials from school methods, school assessment tests, child literature, comic books, and media. The latter category encompasses subtitles of popular TV-programs; RSS feeds from the *Jeugdjournaal* website, a popular youth-oriented daily news broadcast on Dutch television; and texts from websites frequented by children. Roughly, about 42% of the corpus consists of school materials, 38% child literature and comic books, and 20% media. In Table 1 an overview is given of the number of tokens and lemmas in the various domains and over the school grades.

For nearly all materials, the grade indication is based on information provided by the publishers (either regarding grade or age). In the few cases where those were lacking, we estimated the intended grade. Children enter the Dutch elementary school system at age four, in 'Group 1'. Group 1 and 2 are the Kindergarten groups. Formal reading and arithmetics instruction usually start in Group 3, yet much preparatory instruction in these subjects already takes place in Group 2. In Table 1, the internationally more customary Kindergarten and grade indications are used. It should be noted that the totals in the lowest row of Table 1 are not summed totals of the cells above them, since lemmas can overlap in two cells and the totals refer to the number of *different* lemmas. For instance, comics in Grade 4 contain 6,632 different lemmas while comics in Grade 5 contain 15,456 different lemmas, yet the lemmas in comics for these two grades are overlapping. Similarly, whereas the total of tokens at the right (given in bold) is the sum of the column as well as of the row, the total of lemmas, also given in bold, is not the sum of the column and the row but simply represents the

Table 1: BasiLex: overview of the number of tokens and lemmas in the various domains and over school grades

| **TOKENS** | children's literature | comics | school texts: language | school texts: arithmetics | school texts: sciences | school tests | subtitles | newsfeeds | Total |
|---|---|---|---|---|---|---|---|---|---|
| Preschool, K1&2 | 174,228 | 3,700 | | | | | | | 177,928 |
| Grade 1 | 285,699 | 14,612 | 14,171 | | | 9,871 | | | 324,353 |
| Grade 2 | 1,493,478 | | 122,422 | 14,065 | 6,231 | 51,415 | | | 1,687,611 |
| Grade 3 | 241,809 | | 556,727 | 58,397 | 349,959 | 73,812 | 999,255 | | 2,279,959 |
| Grade 4 | 206,010 | 24,327 | 943,750 | 54,489 | 450,495 | 95,826 | 60,395 | 1,168,047 | 3,003,339 |
| Grade 5 | 1,775,384 | 183,328 | 1,037,474 | 24,683 | 521,755 | 105,215 | 144,765 | | 3,792,604 |
| Grade 6 | 172,090 | | 744,610 | 32,916 | 504,460 | 198,797 | 38,409 | | 1,691,282 |
| Sec.sch. 1&2 | 1,030,655 | | | | | | | | 1,030,655 |
| Total | 5,379,353 | 225,967 | 3,419,154 | 184,550 | 1,832,900 | 534,936 | 1,242,824 | 1,168,047 | ***13,987,731*** |

| **LEMMAS** | children's literature | comics | school texts: language | school texts: arithmetics | school texts: sciences | school tests | subtitles | newsfeeds | Total |
|---|---|---|---|---|---|---|---|---|---|
| Pre-school, K1&2 | 6,928 | 413 | | | | | | | 7,006 |
| Grade 1 | 5,119 | 812 | 1,007 | | | 1,225 | | | 9,515 |
| Grade 2 | 26,860 | | 6,632 | 1,216 | 894 | 3,587 | | | 29,732 |
| Grade 3 | 9,444 | | 17,562 | 3,704 | 10,430 | 5,790 | 29,184 | | 48,665 |
| Grade 4 | 10,388 | 2,732 | 27,632 | 4,345 | 14,501 | 8,700 | 5,068 | 32,351 | 62,923 |
| Grade 5 | 32,067 | 15,456 | 32,986 | 3,917 | 16,804 | 9,777 | 7,268 | | 73,261 |
| Grade 6 | 9,974 | | 28,263 | 3,830 | 16,509 | 13,176 | 3,534 | | 47,144 |
| Sec.sch. 1&2 | 27,563 | | | | | | | | 27,563 |
| Total | 68,326 | 16,857 | 65,361 | 10,919 | 34,060 | 23,052 | 33,338 | 32,351 | ***168,073*** |

total of lemmas in BasiLex. Furthermore, the grade indications must be seen as starting points of ranges. For example, the 1,168,047 newsfeeds tokens are tokens in newsfeeds meant to be read by children from Grade 4 *onwards*. Finally, the tokens and lemmas referred to as children's literature and comics for preschool, K1, K2 (i.e., preschool and Kindergarten 1 and 2, Group 1 and 2 in the Dutch system) are books and comics that are either meant to be read to children or are picture books with named pictures, usually also meant to be read by adults and children together.

The collected materials were transformed to the FoLiA format[1] (Format for Linguistic Annotation, Van Gompel and Reynaert 2013, Van Gompel 2014), an XML-based format. Thereupon, the linguistic analysis was performed with Frog,[2] a tokenizer, POS-tagger, lemmatizer and morphological segmenter of word tokens in Dutch text files (Van den Bosch et al. 2007). To each sentence in a text, Frog also adds a dependency graph; the base phrase chunks in the sentence are identified; and all named entities, as far as possible, are identified and labeled as person, organization, location, or miscellaneous. The texts were subsequently cut into paragraphs and these paragraphs are accessible in the corpus in randomized order only, which was a requirement of the publishers who handed over their materials. We defined a paragraph as a text fragment between two indented lines, or between two extra spaces between the lines. In texts where there were no indentations or extra spaces – often texts for younger children – we made cuts after each ten lines.

All uniquely occurring lemmas in BasiLex were collected in a list and tagged for several different word properties that frequently are researched or used as control variables, particularly in psycholinguistic and linguistic research and in language assessment test development. The following word property tags were added: bigram and trigram frequency, corpus frequency, corpus dispersion, family size and family frequency, orthographic neighborhood size and orthographic neighborhood

1. http://proycon.github.io/folia/
2. http://ilk.uvt.nl/frog

frequency, and word length (for a discussion of these word properties see Wauters et al. 2003, Balota et al. 2006, Perdijk et al. 2012).

In view of the construction of a target list of words that children should know at the end of elementary school, we selected the 20,000 most frequent words from the BasiLex lexicon for lexical semantic disambiguation. In this set, there were 6,993 ambiguous words, they had more than one meaning, and 3,881 of these had more than two senses according to the Cornetto lexical-semantic database (Vossen et al. 2013). Disambiguation was done partly automatically (2,972 words) and partly by hand (2,212 words), with the Semantic Annotation Tool developed in the DutchSemCor-project (Vossen et al. 2011), with which senses from the Cornetto database are assigned to disambiguated words. Thus, 5,184 words were disambiguated. The remaining 1,809 words we hope to disambiguate in the near future. The meanings assigned to the disambiguated words refer to Cornetto-senses.

The validity and representativity of BasiLex was checked by comparison to a much-used Dutch word list for children and by checking on words in specific domains such as words for weekdays and months. Furthermore, we analyzed frequency distributions in BasiLex.

The *Streeflijst woordenschat voor 6-jarigen* ("Target list vocabulary knowledge for 6-year-olds", Schaerlaekens et al. 1999) is a word list, partly based on earlier composed word lists, among which Schrooten and Vermeer (1994). For the *Streeflijst* each word was evaluated by professionals who were familiar with the vocabulary of six-year old Dutch and Flemish L1-children. Both graduated and nearly graduated professionals selected and judged the words. The list of 6,380 partly disambiguated words that resulted was then presented to 182 Dutch and Flemish teachers, half of which worked in the last year of Kindergarten and the other half in Grade 1. They were asked to indicate for each word whether they thought a six-year old ought to know it. For 1,536 of the 6,380 words, 90% of these 182 persons indicated that a six-year-old should "know and understand" them. This 1,536-word list by Schaerlaekens et al. is known as the *Unaniemenlijst* ("unanimity list"), and this list we used for a comparison with BasiLex.

All 1,536 words were present in the BasiLex lexicon of single words, except for 13 that were constructions rather than single words (e.g., *een voor een*, one by one), 11 that in the *Streeflijst* were inflections or conjugations instead of lemmas, and one other word (*rolschaatsen*, roller skating, possibly because this is hardly done by Dutch children nowadays). Of these 1,536, 21.3% words are among the 500 most frequent BasiLex lemmas; 81.7% are among the 5,000 most frequent BasiLex lemmas.

Another way to get an impression of the validity and representativity of BasiLex is to compare the relative frequencies of domain-specific draws on, for instance, words for weekdays, months, seasons, wind directions and colors. In a valid and representative corpus all names of weekdays should occur more or less with the same frequency, and similarly all names of months. Colors are expected to occur in a Zipfian-Mandelbrotian distribution (Piantadosi to appear), with red, black, and white occurring most frequently. In Table 2 are given the relative frequencies for some such domain words, for the English translations (NB in Dutch there are two words for 'spring', namely *lente* and *voorjaar*, hence we give the relative frequencies for both).

Table 2 shows that frequencies for words belonging to the same domain indeed are quite similar, and that the frequencies of the color words indeed follow a Zipfian-Mandelbrotian distribution.

These indicators suggest that the BasiLex corpus is a reliable and representative corpus to serve for educational purposes as well as for research regarding children's spoken and written language development.

## 3. Comparing BasiLex with a SoNaR subcorpus

SoNaR is a corpus of Dutch written language comprising over 500 million words from different domains and genres, both from the traditional printed media and from several new, digital media.[3]

3. See http://tst-centrale.org/nl/producten/corpora/sonar-corpus/6-85 for user documentation.

Table 2: Relative frequencies in Basilex for specific domain words

| | | | | | |
|---|---|---|---|---|---|
| rood.red | 0.00033901 | zomer.summer | 0.00010852 | mei.may | 0.0000336 |
| zwart.black | 0.00020068 | winter.winter | 0.00010123 | januari.january | 0.00003024 |
| wit.white | 0.00019474 | lente, | 0.00002467 | oktober.october | 0.00002788 |
| blauw.blue | 0.00017673 | voorjaar.spring | 0.00001523 | december.december | 0.0000271 |
| groen.green | 0.00016672 | herfst.autumn | 0.00002166 | juni.june | 0.00002674 |
| geel.yellow | 0.00010381 | noord.north | 0.00005591 | april.april | 0.00002581 |
| bruin.brown | 0.00007757 | zuid.south | 0.00005262 | juli.july | 0.00002466 |
| lila.lilac | 0.00000114 | west.west | 0.00003088 | november.november | 0.00002374 |
| kobaltblauw.cobalt blue | 0.00000007 | oost.east | 0.00002881 | maart.march | 0.00002266 |
| | | | | september.september | 0.00002109 |
| | | | | februari.february | 0.00001852 |
| | | | | augustus.august | 0.00001844 |

From the SoNaR sub-corpora (Oostdijk et al. 2013) we selected those that were most comparable to the BasiLex sub-corpora, namely: printed books, brochures, newsletters, newspapers and periodicals, and as new media: subtitles, teletext pages, and websites (including Wikipedia). We left out discussion lists, press releases, blogs, tweets (all published and electronic); guides and manuals, legal texts, policy documents, proceedings, reports (all published and printed); chats, sms, written assignments, autocues, texts for the visibly impaired (all unpublished and electronic). This resulted in a sub-corpus of 387,380,713 tokens in which 4,143,383 unique lemmas occur. Figure 1 shows the Zipf scatter plots for all token occurrences in our SoNaR selection and the entire BasiLex corpus, in log-log space.

The two plots approximate the typical Zipf-Mandelbrot curve, confirming that BasiLex (the smaller corpus, hence the lower plot) is as normal a corpus as SoNaR. Yet, Figure 1 reveals nothing about possible differences in frequencies of PoS-classes or of particular words.

To shed more light on such possible differences we compared the frequencies of some of the PoS-classes for the two corpora. We did not do this for the full corpora but for the most frequent 17,000 lemmas in both of them. The figure of 17,000 is related to the number of words children know at the end of elementary school (Kuiken and Vermeer 2013, see also Table 2). Dutch L1 children know about 5,000 words at the age of seven, that is, when they are in Grade 1, and they know about 17,000 words in Grade 6, right before secondary school. The latter figure applies to L1 Dutch children; for L2 Dutch children estimations rank from about 12,000 to 15,000. Of course, although word frequency is the best predictor of word knowledge (Verhoeven et al. 2011, Lee 2011), the words individual children know at the beginning and end of elementary school, respectively, will only partly overlap with the 5,000 and 17,000 most frequent lemmas in BasiLex. Not only because BasiLex does not cover children's actual language experience but also because children, evidently, have enormous amount of language input via oral language. Frequencies of words in oral and written language can differ considerably. Nevertheless, as a starting point for comparing written words children will encounter in elementary school with adult corpora of written language, the 17,000 most frequent lemmas would seem to constitute a viable set. Overall, the BasiLex corpus contains 13,979,214 tokens, in which 168,111 unique lemmas occur. These include punctuation marks, numbers as numerals (so, e.g. 17 for seventeen), non-letter/figure signs (i.e. #, %) and non-words. We removed the latter categories. With the 17,000$^{\text{th}}$ lemma in the frequency list having a frequency of 15, selecting all lemmas with a frequency of higher than 14 resulted in 17,495 lemmas. Thus, we also selected the 17,495 most frequent lemmas in the SoNaR subcorpus; the 17,495$^{\text{th}}$ lemma occurs 863 times in the 387,380,713-token subcorpus of SoNaR. Since the BasiLex lexicon and the SoNaR derived lexicon differ in size, we computed relative instead of absolute frequencies to compare both lexicons. The relative frequency of a lemma was computed by dividing its absolute frequency by the total frequency of all the lemmas in the lexicon. Next, we converted these relative

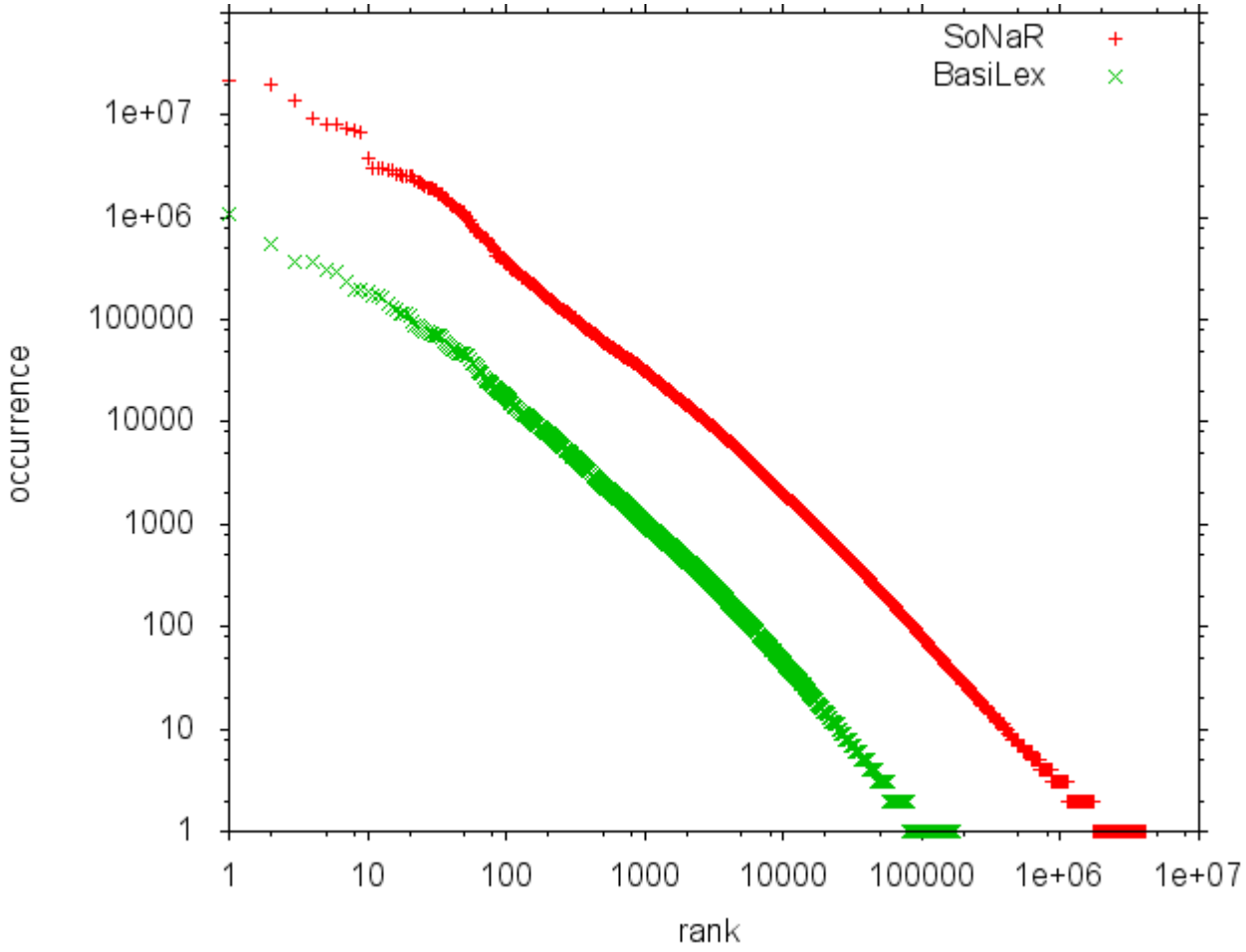


Figure 1: Scatter plots of frequency rank and number of occurrences, in log-log space, for all tokens occurring in the SoNaR sub-corpus and the Basilex corpus

frequencies to log-relative frequencies by taking their base-10 logarithms for more optimal further computations.

### 3.1 Comparing the log-relative frequencies for part-of-speech tags in Basilex and SoNaR

Words in BasiLex and in SoNaR were PoS-tagged by Frog with tags for adjectives, adverbs, articles, conjunctions, interjections, nouns, verbs, prepositions, pronouns, and 'specials' (SPECs, see below). With a Chi-square test we analyzed whether the distribution of log-relative frequencies for the tags in the two 17,495 word sets differed. Table 3 presents the distribution of part-of-speech tags in the two corpora.

The Chi-square test showed, not surprisingly, given the small differences in percentages, that the distributions did not differ significantly ($\chi^2 = 2.2$, p= 0.0001, which is very far from the critical value of 23.21 for 10 degrees of freedom, see Field (2009)).

We then performed Independent Samples t-tests on all PoS tags except SPEC. Next to first and last names, the SPEC label covers words referring to subject matters (geography, history, languages),

Table 3: Distribution of frequencies of PoS tags in BasiLex and SoNaR (in percentages)

| | N | V | Adj | Adv | Cnj | Prep | Prn | Ijct | Num | Art | SPEC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BasiLex | 46.4 | 15.1 | 9.8 | 2.3 | 0.2 | 0.3 | 0.5 | 0.2 | 0.4 | 0.03 | 25.0 |
| SoNaR | 40.5 | 14.9 | 12.8 | 2.5 | 0.3 | 0.4 | 0.6 | 0.1 | 0.4 | 0.1 | 27.5 |

titles of films and TV-programs, locations, organizations (including acronyms), and words from other languages (e.g. *eau*). One reason we left the SPECs out is that most of them are not very relevant for the BasiLex-SoNaR comparison (e.g. last names, titles of films). Another one is that using them for comparisons requires fine-tuned analyses that go beyond the framework of this article (e.g. *Willem* might have a certain frequency in SoNaR as a Dutch first name like many other first names, yet in BasiLex it might have a certain frequency as part of *Willem I*, *Willem II*, and *Willem III*, as these appear as former kings of the Netherlands in educational history texts).

In the BasiLex set there were 4,253 SPECs, so 13,242 lemmas remained. In the SoNaR set there were 4,638 SPECs, so 12,857 lemmas remained. We removed an additional four lemmas from SoNaR and nine from BasiLex as incorrectly tagged lemmas. We calculated log-relative frequencies for all of them and then performed Independent Samples t-tests with Corpus as grouping variable. For those PoS tags where Levene's tests showed that equality of variance could not be assumed, we give the results for the t-test for unequal variances as well as the results of an additional non-parametric test, namely, Kruskal-Wallis. Where Levene's tests showed that equal variances could be assumed, extra non-parametric tests were not necessary. The results are given in Table 4.

Table 4: Results of t-tests and non-parametric tests for PoS's in BasiLex versus SoNaR

| | SoNaR lemmas | M log.freq. SoNaR | SD | BasiLex Lemmas | M log.freq. BasiLex | SD | Equal var? | t-test results | Kruskal-Walis results |
|---|---|---|---|---|---|---|---|---|---|
| N | 7,079 | -5.11 | 0.476 | 8,067 | -5.285 | 0.541 | no | .000 (t=-21.184) | .000 |
| V | 2,653 | -5.021 | 0.542 | 2,690 | -5.161 | 0.637 | no | .000 (t=-8.677) | .000 |
| Adj | 2,262 | -5.053 | 0.514 | 1,724 | -5.202 | 0.622 | no | .000 (t=-8.268) | .000 |
| Adv | 469 | -4.723 | 0.714 | 431 | -4.802 | 0.82 | no | n.s. | .014 |
| Cnj | 54 | -4.389 | 0.982 | 36 | -4.225 | 1.148 | yes | n.s. | |
| Prp | 92 | -4.09 | 1.116 | 63 | -3.887 | 1.205 | yes | n.s. | |
| Prn | 134 | -4.124 | 1.006 | 106 | -4.008 | 1.226 | no | n.s. | n.s. |
| Ijct | 18 | -4.942 | 0.684 | 28 | -4.939 | 0.78 | yes | n.s. | |
| Num | 79 | -4.612 | 0.748 | 76 | -4.927 | 0.8 | yes | .012 (t=-2.534) | |
| Art | 12 | -3.732 | 1.327 | 7 | -3.4633 | 1.847 | yes | n.s. | |

It should be remembered that logarithms for numbers between zero and one – which relative frequencies are – are negative; and that the more negative the logarithm of a number is, the closer to zero that number is. Thus, the significant differences found for the nouns, verbs, adjectives, adverbs (only in the Kruskal-Wallis) and numerals all imply that the relative frequency of those PoS tags is mildly higher in SoNaR than in Basilex. For all other PoS tags – all of them function[4] words – no significant differences in relative log-frequencies were found for BasiLex versus SoNaR. We calculated effect sizes (i.e., Cohen's *d*'s) for the significant differences; they are 0.34 for the Nouns, 0.24 for the Verbs, 0.26 for the Adjectives, and 0.40 for the Numerals, which are all small effect sizes.

4. Throughout this paper, *content words* refers to adjectives, adverbs, nouns, verbs and interjections. *Function words* are all other parts of speech, except for SPECs, which fall outside these categories. See e.g. Finegan (1994), p.161.

### 3.2 Frequency profiling for BasiLex and SoNaR

We performed frequency profiling as described in Rayson and Garside (2000), who developed a log-likelihood metric to compare corpora, accounting both for the sizes of the compared corpora and for the frequency ranks of the words. The expected value for each target word in each of both corpora is calculated, relative to the size of the corpus, and then the log-likelihood of the word over the two corpora is calculated. This results in a word list with log-likelihood values (LL values), in which the word with the largest LL value has the largest relative frequency difference over both corpora. With this formula, in which frequencies are multiplied with frequency differences, frequency differences between words with high frequencies are more important than frequency differences between words with low frequencies. This is what one would wish, since a large relative frequency difference between words that have low frequencies will usually be viewed as being less important than a smaller relative frequency difference between words that are highly frequent. Aside from computing log-likelihoods we also performed t-tests and computed mutual information, which as expected were too sensitive to high-frequency words and low-frequency words, respectively; in line with Kilgarriff (2001) we opted for comparing log-likelihoods between the frequencies in our two differently sized corpora.

Also for the frequency profiling, from the 17,495 BasiLex and SoNaR lemmas we left out those tagged as SPEC (specials). There were 8,854 lemmas that occurred in both sets and, thus, could be used for the frequency profiling. We first inspected the 500 words with the highest LL value and the 500 with the lowest LL value. The former contained 111 function words (17,7% of all function words in the total set) while the latter contained 26 (4,2%). We then split up the list of 8,854 LL values into function words (total N = 625) and content words (total N=8,229) and inspected the 500 content words with the highest LL values, that is, for which the two corpora differ the most.

Among the 500 content words with the highest LL values are many typical school words, such as: *woord* (word), *zeggen* (to say), *schrijven* (to write), *zin* (sentence), *lezen* (to read), *juf* and *meester* (teacher, the female and the male variant, respectively), *opgave* (question/assignment), *tekst* (text), *bedenken* (to think about). All these are in the top-25 of LL-values and having higher relative frequencies in BasiLex. Among the adjectives, there are several words that would be expected to occur more in children's literature than in adult literature: *stom* (stupid), *slim* (smart), *raar* (weird), *gek* (crazy), which indeed have higher relative frequencies in BasiLex. Higher relative frequencies in SoNaR have words one would expect to occur more often in adult language: *bedrijf* (company), *politiek* (political), *sociaal* (social), *Europees* (European), *economisch* (economic). The nouns with higher relative frequencies in BasiLex contain several words for family members, *opa* (gramps, grandfather), *oma* (granny, grandmother), *vader* (father), *moeder* (mother), *oom* (uncle), *mam* (mom), *pap* (dad), *mama* (mama) whereas SoNaR has higher relative frequencies for words having to do with governing, *gemeente* (municipality), *overheid* (government), *voorzitter* (chairman), *politiek* (politics). As regards the verbs, next to the typical school verbs already mentioned, there are verbs like to *giechelen* (to giggle), *zwemmen* (to swim), *fietsen* (to cycle) versus *organiseren* (to organize), *bevestigen* (to affirm), *verklaren* (to declare). Little surprising as these differences between BasiLex and SoNar might seem, they do confirm the validity of BasiLex as representative of child-directed writing. In the Appendix we give the one hundred content words with the highest LL-values and their relative frequencies in both corpora.

## 4. Comparing BasiLex with Schrooten and Vermeer

Also for this analysis we computed relative frequencies (in the same manner as for the BasiLex SoNaR comparison, see above) and log-relative frequencies. Schrooten and Vermeer (1994) (S&V, for short) is a smaller lexicon than BasiLex; its 17,495$^{\text{th}}$ lemma has a frequency of 1. S&V does not contain the type of words tagged as SPEC so there was no need to remove these. Furthermore, in contrast to BasiLex and SoNaR, the lemmas in S&V have no PoS tags, so a comparison regarding part of speech, as we did for BasiLex and SoNaR, is not possible here. The ambiguous words are

disambiguated (manually). Yet whereas in BasiLex and SoNaR a letter string has more than one entry only when it has different PoS tags (e.g., *aanzien* as noun and as verb are two entries), in S&V each and every meaning has a separate entry with a separate frequency tag. Thus, there are, for instance, 28 entries for *gaan* (to go). Therefore, we transformed the S&V word set into a lemma list with summed frequencies (e.g. the 28 frequencies for *gaan* were summed into one frequency for the one entry *gaan*). This reduced the S&V word set to 11,562 lemmas. We performed a similar operation on the BasiLex entries, that is for identical letter strings with different PoS tags in the list we summed the frequencies (e.g. the verb *aanzien* and the noun *aanzien*), resulting in a list of 14,023 lemmas for BasiLex without PoS tags. Of the 11,562 S&V-lemmas, 8,225 were in the BasiLex list. We then tagged these lemmas as being either a function word or a content word. For letter strings that could be both (e.g., *elf* fairy/eleven), we gave the tag that occurred most frequently in BasiLex. Thus we found 7,668 content words and 537 function words. With these we performed Rayson & Garside's frequency profiling.

We inspected the 500 words with the highest LL value. In contrast to the comparison with SoNaR, these contained many function words (100 out of 536) and many auxiliaries. For instance, the verbs *zijn* (to be), *hebben* (to have), *gaan* (to go), *kunnen* (to be able to), *moeten* (to have to, must), *zullen* (shall), *mogen* (may, can, should), *willen* (to want) were among the 30 words with the highest LL values. All of these had a higher relative frequency in S&V than in BasiLex. In general, most of the function words have a higher relative frequency in S&V than in BasiLex. One possible explanation for this difference lies in the nature of the materials in both corpora. Subtitles, which are part of BasiLex yet not of S&V, usually contain fewer auxiliaries because they are made as short as possible. Also, BasiLex contains educational training materials, which S&V does not contain; these also might contain fewer auxiliaries. However, more detailed analyses are needed to explain the relative frequency differences regarding auxiliaries and function words in BasiLex and S&V.

Apart from auxiliaries and function words, three categories of words among the one hundred words with the highest LL-values are noticeable. First, words with spelling differences between 1994 and 2014, as in *paddestoel* and *paddenstoel* (mushroom); second, words for objects that are not longer current or just have become current, like *gulden* (guilder) versus *euro*. Third, as S&V also included spoken language addressed to children (teachers' input in the classroom, 540,000 tokens), a dozen of words referring to classroom interactions are relatively more frequent in S&V, such as *hé* (hey), *ophouden* (to stop), *verdergaan* (to go on), *nakijken* (to check/to mark). The Appendix lists the one hundred content words with the highest LL-values for BasiLex and S&V and their relative frequencies.

## 5. BasiLex and possible educational applications

The BasiLex corpus gives insight into the sort of written words and sentences primary school children typically will encounter, with what characteristics, from what source, in what environment, at what age, and how often. This will help educational advisors, curriculum designers, and teachers to estimate the prior word knowledge of children from different ages and different backgrounds. Knowing words is the key to understanding and being understood. One of the major obstacles preventing children from doing well at school seems to be their limited lexical abilities: lexical skills correlate highly with other language abilities, for instance, reading abilities (Verhoeven and Vermeer 2006). Reading is a central activity in school, not only in language classes, but also in other subjects. Children with a limited knowledge of words will have serious problems comprehending texts.

Words are not learned in an arbitrary order. Huttenlocher et al. (1991) found that the relative frequency of words from the parents' input is strongly related to the order of acquisition of those words by their children. High-frequency words are better known than low-frequency ones (Brown 1993, Vermeer 2001). High-frequency words in texts are recognized faster (Rudell 1993); for more

experimental results on the effects of word frequency and age of acquisition in recognition and recall, see also Dewhurst et al. (1998), and Gerhand and Barry (1998).

### 5.1 Word selection for curriculum materials

Since the frequency of a word is related to acquisition order, the word frequencies in BasiLex can be used as a criterion to select words as target words for a specific age group or grade in primary school. In order to facilitate the choice of words from the BasiLex corpus for teachers and curriculum designers, we distinguished eleven categories of frequency classes (we will call them vocabulary lists, henceforth *voclists*) based on the order of the word frequencies in BasiLex, see Table 5 below. The first voclist consists of the 1,000 lemmas that have the highest frequency, in other words, the thousand most frequent words in daily written input in elementary school. The second voclist consists of the 1,000 lemmas that follow, et cetera. Voclists 6, 7 and 8 consist of 1,500, 2,000 and 2,500 lemmas, respectively; voclist 9 and 10 have 4,000 and 5,000 lemmas, respectively, and voclist 11 contains the tail of the Zipfian distribution with over 80,000 lemmas. The first 1,000 lemmas in the first voclist account for 82.7% of the tokens in the entire corpus, as can be seen in the fourth column in Table 5. The second voclist with 1,000 lemmas covers 5.3% of the corpus. In line with Zipf's law (cf. Figure 1), whereas the 80,000 least frequent words account for less than 3% of all tokens in the corpus. Table 5 shows in the fifth column the frequency bands of the voclists. As can be seen, the 20,000 most frequent lemmas in BasiLex have a frequency of occurrence of 9 or more.

In the last columns, the mean size of receptive word knowledge per age group is given (as indicated by Kuiken and Vermeer 2013). On the basis of these figures, an indication is given of word selection as target words for a specific grade in elementary school (see above for an explanation of the Dutch elementary school system). For example, for children in Grade 1 (6 year olds), who have a mean size of receptive vocabulary of about 4,500 lemmas, words can be chosen as target words from voclists 3, 4 and 5 (the first 5,000 words), having a frequency of occurrence in BasiLex between 380 and 99. For a more elaborate description of such a word selection procedure for educational purposes, see Van de Guchte and Vermeer (2003).

Table 5: Word lists in BasiLex: number of lemmas, cumulative number of lemmas, token coverage, frequency bands, and word selection criteria, on each frequency level/voclist.

| frequency levels | lemmas in voclist | N lemmas, cumulative | % token coverage | frequency band BasiLex | word selection per grade | receptive word knowl. per age gr. |
|---|---|---|---|---|---|---|
| voclist 1 | 1,000 | 1,000 | 82.7 | >922 | KG 1 and 2 | |
| voclist 2 | 1,000 | 2,000 | 5.3 | 922 to 381 | KG 1 and 2 | |
| voclist 3 | 1,000 | 3,000 | 2.6 | 380 to 215 | KG to 1 | 4: 3,000 |
| voclist 4 | 1,000 | 4,000 | 1.6 | 214 to 139 | KG to 2 | 5: 3,800 |
| voclist 5 | 1,000 | 5,000 | 1.1 | 138 to 99 | 1 to 3 | 6: 4,500 |
| voclist 6 | 1,500 | 6,500 | 1.1 | 98 to 60 | 2 to 4 | 7: 5,200 |
| voclist 7 | 2,000 | 8,500 | 0.9 | 59 to 41 | 3 to 5 | 8: 6,000 |
| voclist 8 | 2,500 | 11,000 | 0.7 | 40 to 26 | 4 to 6 | 9: 8,500 |
| voclist 9 | 4,000 | 15,000 | 0.7 | 25 to 15 | 5 and 6 | 10: 11,000 |
| voclist 10 | 5,000 | 20,000 | 0.5 | 14 to 9 | 5 and 6 | 11: 14,000 |
| voclist 11 | >80,000 | 110,000 | 2.8 | 8 or lower | 6 | 12: 17,000 |

### 5.2 Measuring vocabulary size and text difficulty

Since the frequency of a word is related to acquisition order, a valid measure of the lexical difficulty of a text might be to relate the words in that text to their frequency (or frequency classes, voclists) in a reference corpus such as BasiLex. This is comparable to a procedure in the Lexical Frequency Profile (LFP, Laufer and Nation 1995) for written texts, in which four levels are distinguished, or the

measure for Advanced Lexical Richness (ALR, Treffers-Daller and Van Hout 1999, Daller et al. 2003) for spoken texts, in which two frequency levels are distinguished. Moreover, a measure of lexical difficulty based on a reference frequency list allows to give an indication of the absolute vocabulary size of a writer/speaker, in the same way as extrapolating scores on some vocabulary tests (cf. the Vocabulary Levels Test, Laufer and Nation 1999). Such a procedure for calculating text difficulty and indication of vocabulary size is performed to compute the Measure of Lexical Richness (MLR Vermeer 2004b, Vermeer 2004a) with Schrooten and Vermeer (1994) as reference corpus.

One way of indicating the level of difficulty of a particular written text is to calculate the relative number of known versus unknown tokens or lemmas in a text ('text coverage'; see, e.g. Carver 1994, Hsueh-Chao and Nation 2000). The percentage of text coverage can indicate the degree of difficulty of a text for a reader. The precise relationship between text coverage and text comprehension is dependent on various factors, such as the reader's world knowledge, the subject matter, the number of cognates, the style, the text type. For reasonable comprehension of a text, a token coverage of 95-96% is considered to be a threshold, or a lower lemma coverage of about 87% (cf. Hazenberg 1994). In Goossens and Vermeer (2009), 175 pupils of Grade 4 were divided into four groups with different vocabulary sizes, on the basis of a vocabulary test. Next, they read six texts differing in word-level difficulty. Each time after they had read a text, the children answered questions about it, and their knowledge of unknown words from the text was tested. The results show that the more low-frequency words there were in the texts, the less well the children understood the texts, and the fewer words they learned from them. Optimal text coverage was reached at a level of 88 percent, with the children answering correctly more than half of the text comprehension questions, and having learned at least one out of five unknown words from the text.

So, combining two indicators, text difficulty ('lemma coverage') and user proficiency ('vocabulary size'), leads to an assessment of the appropriateness of a given text for a given user. The number of lemmas assessed as being known to a learner (e.g., by extrapolating scores on a vocabulary test for a specific learner, or on the basis of the mean receptive knowledge of a specific age group) can be calculated for each individual text in a text coverage percentage. For example, six-year-olds in Grade 1 have a receptive vocabulary of 4,500 lemmas (see Table 5). By calculating in a particular text the relative number of known words (the 5,000 lemmas from the first five voclists) divided by the total number of lemmas in that text, it is known whether that text is comprehensible to Grade 1 children (having a lemma coverage of around 88%). In a web application of BasiLex we will build in the near future, it will be possible to upload a text to find out what the percentage text coverage of that text is for various vocabulary levels.

In a comparable way, the wording in a particular text of a writer (or speaker) can give an indication of the productive vocabulary size of that writer/speaker. Like an extrapolated score on a vocabulary test related to a dictionary, the words used by that writer/speaker related to the BasiLex corpus can give an indication of a person's productive vocabulary size. To calculate such an indication, the relative distribution of the token coverage in the BasiLex corpus in the fourth column in Table 5 can be taken as a model. In other words, if the relative distribution of the words of an analyzed text over the eleven voclists is the same as those in the fourth column in Table 5, then the indicated vocabulary size is considered to match with a vocabulary of about 100,000 words. If a person uses relatively more words from the first four voclists, then his score is lower. If someone uses words from the first voclist only, his vocabulary score is 1 (indicating a vocabulary size of about 1,000 words). The vocabulary score is thus calculated by adding up each quotient of the text coverage of the text of the writer (or transcript of the speaker), and the 'model' coverage of BasiLex ('token coverage' in Table 2) of each voclist.

## 6. Conclusions

We presented BasiLex, a 13.5 million-token, 11.5-million words corpus of texts written for children in Kindergarten and primary school. The corpus consists of school materials (42%), child literature and

comics (38%), and media (20%; subtitles, child-oriented news items, and texts from child-oriented websites). The corpus fills a need: the most comparable corpus, by Schrooten and Vermeer (1994), is smaller and outdated. A comparison with the Schrooten and Vermeer corpus indeed shows high log-likelihood scores for time markers such as *euro* versus *guilder*. A comparison with the current largest corpus of contemporary written Dutch for adults, SoNaR (Oostdijk et al. 2013), shows that the largest frequency deviations occur with content words typical of children's versus adult domains. In contrast, we find no significant deviations in the relative frequencies of part-of-speech tags.

Adopting the FoLiA XML format (Van Gompel and Reynaert 2013, Van Gompel 2014), the corpus is automatically annotated for part-of-speech, lemma, and other linguistic annotation layers provided by Frog (Van den Bosch et al. 2007). Polysemous words in the top-20,000 of most frequent lemmas have been annotated partially automatically and partly manually with Cornetto word senses (Vossen et al. 2013). Automatic tagging comes with a margin of error; while part-of-speech tagging on main tags is estimated to produce errors in only 1.4% of all tokens (15.7% error on unseen words, Van den Bosch et al. 2007), automatic sense tagging is considerably less reliable (Vossen et al. 2011).

We have exemplified how BasiLex could be used for educational applications and diagnostics for texts. We aim to use the BasiLex corpus and lexicon in conjunction with its sister project BasiScript, a corpus of texts produced by children of the same ages (partially longitudinally collected), to be finished in the second half of 2015, in further studies in which we aim to relate the frequency of occurrence of both single words and $n$-grams in the two corpora.

BasiLex will be available in the course of 2014 via the TST Centrale. We express the hope that the resource will be used in academic studies in a wide range of research areas, as well as in the educational domain.

## References


Arnon, I. and N. Snider (2010), More than words: Frequency effects for multi-word phrases, *Journal of Memory and Language* **62** (1), pp. 67–82, Elsevier.

Baayen, R.H., R. Piepenbrock, and H. Van Rijn (1993), The CELEX lexical database on cd-rom, *The Linguistic Data Consortium*, Philadelphia, PA.

Bacchini, S., T. Boland, M. Hulsbeek, and M. Smits (2005), *Duizend-en-één-woorden. De allereerste Nederlandse woorden voor anderstalige peuters en kleuters*, Stichting Leerplanontwikkeling, Enschede, The Netherlands. Available at: http://catalogus.slo.nl.

Balota, D.A., M.J. Yap, and M.J. Cortese (2006), Visual word recognition: The journey from features to meaning, *Handbook of psycholinguistics* **2**, pp. 285–375, Academic Press.

Bannard, C. and D. Matthews (2008), Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations, *Psychological Science* **19** (3), pp. 241–248, SAGE Publications.

Brown, C. (1993), Factors affecting the acquisition of vocabulary: Frequency and saliency of words, *in* Th. Huckin, M. Haynes and J. Coady, editors, *Second language reading and vocabulary learning*, pp. 263–286.

Carver, R.P. (1994), Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction, *Journal of Literacy Research* **26** (4), pp. 413–437, SAGE Publications.

Coenen, M. and A. Vermeer (1988), *Nederlandse woordenschat allochtone kinderen*, Zwijsen, Tilburg, The Netherlands.

Cucchiarini, C., H. Van Hamme, O. Van Herwijnen, and F. Smits (2006), JASMIN-CGN: Extension of the spoken Dutch corpus with speech of elderly people, children and nonnatives in the human-machine interaction modality, *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, pp. 135–138.

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

De la Court, J.F.H.A. (1937), *De meest voorkomende woorden en woordcombinaties in het Nederlandsch: verslag van een onderzoek in opdracht van het Departement van Onderwijs en Eeredienst*, Batavia: Volkslectuur.

Dewhurst, S.A., G.J. Hitch, and C. Barry (1998), Separate effects of word frequency and age of acquisition in recognition and recall, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24** (2), pp. 284–298, American Psychological Association.

Elman, J.L. (2009), On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon, *Cognitive science* **33** (4), pp. 547–582, Wiley Online Library.

Field, A. (2009), *Discovering statistics using SPSS, third edition*, SAGE publications.

Finegan, E. (1994), *Language: Its Structure and Use*, Harcourt Brace College Publishers.

Frank, S.L., R. Bod, and M.H. Christiansen (2012), How hierarchical is language use?, *Proceedings of the Royal Society B: Biological Sciences* **279** (1747), pp. 4522–4531, The Royal Society.

Gerhand, S. and C. Barry (1998), Word frequency effects in oral reading are not merely age-of-acquisition effects in disguise, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24** (2), pp. 267–283, American Psychological Association.

Goossens, N. and A. Vermeer (2009), Wat is een optimale tekstdekking? Woordkennis en tekstbegrip in groep 6, *Toegepaste Taalwetenschap in Artikelen* **82**, pp. 81–92.

Hazenberg, S. (1994), *Een keur van woorden: de wenselijke en feitelijke receptieve woordenschat van anderstalige studenten*, PhD thesis, Vrije Universiteit, Amsterdam, The Netherlands.

Hsueh-Chao, M.H. and P. Nation (2000), Unknown vocabulary density and reading comprehension, *Reading in a foreign language* **13** (1), pp. 403–430.

Huttenlocher, J., W. Haight, A. Bryk, M. Seltzer, and T. Lyons (1991), Early vocabulary growth: Relation to language input and gender, *Developmental psychology* **27** (2), pp. 236–248, American Psychological Association.

Kilgarriff, A. (2001), Comparing corpora, *International journal of corpus linguistics* **6** (1), pp. 97–133.

Kuiken, F. and A. Vermeer (2013), *Nederlands als tweede taal in het basisonderwijs*, ThiemeMeulenhoff, Utrecht, The Netherlands.

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

Laufer, B. and P. Nation (1999), A vocabulary-size test of controlled productive ability, *Language testing* **16** (1), pp. 33–51, SAGE Publications.

Lee, J. (2011), Size matters: Early vocabulary as a predictor of language and literacy competence, *Applied Psycholinguistics* **32** (01), pp. 69–92, Cambridge University Press.

Martin, W. (1967), *De inhoud van krant en roman: een frequentieonderzoek*, Plantijn, Antwerpen, Belgium.

Oostdijk, N., M. Reynaert, V. Hoste, and I. Schuurman (2013), The construction of a 500-million-word reference corpus of contemporary written Dutch, *in* Spyns, P. and J. Odijk, editors, *Essential Speech and Language Technology for Dutch*, Springer, pp. 219–247.

Perdijk, K., R. Schreuder, R.H. Baayen, and L. Verhoeven (2012), Effects of morphological family size for young readers, *British Journal of Developmental Psychology* **30** (3), pp. 432–445, Wiley Online Library.

Piantadosi, S.T. (to appear), Zipfs word frequency law in natural language: A critical review and future directions, *Psychonomic Bulletin & Review*. in press, published online DOI 10.3758/s13423-014-0585-6.

Rayson, P. and R. Garside (2000), Comparing corpora using frequency profiling, *Proceedings of the workshop on Comparing Corpora*, Association for Computational Linguistics, pp. 1–6.

Rudell, A.P. (1993), Frequency of word usage and perceived word difficulty: Ratings of Kučera and Francis words, *Behavior Research Methods, Instruments, & Computers* **25** (4), pp. 455–463, Springer.

Schaerlaekens, A.M., D. Kohnstamm, and M. Lejaegere (1999), *Streeflijst woordenschat voor zesjarigen. Derde herziene versie. Gebaseerd op nieuw onderzoek in Nederland en België*, Swets & Zeitlinger, Lisse, The Netherlands.

Schrooten, W. and A. Vermeer (1994), *Woorden in het basisonderwijs. 15.000 woorden aangeboden aan leerlingen*, Tilburg University Press, Tilburg, The Netherlands. Available at: http://www.woordwerken.annevermeer.com.

Treffers-Daller, J. and R.W.N.M. Van Hout (1999), De meting van woordenschatrijkdom in het Turks van Turks-Duits tweetaligen, *in* Huls, B. and B. Weltens, editors, *Artikelen van de derde sociolinguïstische conferentie*, Eburon, Delft, The Netherlands, pp. 428–440.

Uit den Boogaart, P.C., editor (1975), *Woordfrequenties in geschreven en gesproken Nederlands*, Oosthoek, Scheltema & Holkema, Utrecht, The Netherlands.

Van Berckel, J.A.Th.M. (1962), Onderzoek woordfrequentie: resultaten kranten, *Technical report*, Stichting Mathematisch Centrum, Amsterdam, The Netherlands.

Van de Guchte, C. and A. Vermeer (2003), Een passende woordkeus: Het kiezen van woorden voor woordenschatlessen, *Toegepaste Taalwetenschap in Artikelen* **69** (1), pp. 9–23.

Van den Bosch, A., B. Busser, S. Canisius, and W. Daelemans (2007), An efficient memory-based morphosyntactic tagger and parser for Dutch, *Computational linguistics in the Netherlands: Selected papers from the Seventeenth CLIN Meeting*, pp. 99–114.

Van Gompel, M. (2014), FoLiA: Format for linguistic annotation. documentation, *Technical Report Language and Speech Technology Technical Report Series LST-14-01*, Radboud University Nijmegen.

Van Gompel, M. and M. Reynaert (2013), FoLiA: A practical XML format for linguistic annotation - a descriptive and comparative study, *Computational Linguistics in the Netherlands Journal* **3**, pp. 63–81.

Verhoeven, L. and A. Vermeer (2006), Literacy achievement of children with intellectual disabilities and differing linguistic backgrounds, *Journal of Intellectual Disability Research* **50** (10), pp. 725–738, Wiley Online Library.

Verhoeven, L., J. van Leeuwe, and A. Vermeer (2011), Vocabulary growth and reading development across the elementary school years, *Scientific Studies of Reading* **15** (1), pp. 8–25, Taylor & Francis.

Vermeer, A. (2001), Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input, *Applied Psycholinguistics* **22** (02), pp. 217–234, Cambridge University Press.

Vermeer, A. (2004a), MLR: Maat voor Lexicale Rijkdom/Measure of Lexical Richness. Available at: http://www.woordwerken.annevermeer.com.

Vermeer, A. (2004b), The relation between lexical richness and vocabulary size in Dutch L1 and L2 children, *in* Bogaards, P. and B. Laufer, editors, *Vocabulary in a second language: Selection, acquisition and testing*, Vol. 10, John Benjamins Publishing, Amsterdam/Philadelphia, pp. 173–189.

Vossen, P., A. Görög, F. Laan, M. van Gompel, R. Izquierdo, and A. van den Bosch (2011), DutchSemCor: building a semantically annotated corpus for Dutch, *in* Kosem, I. and K. Kosem, editors, *Electronic lexicography in the 21st century: New applications for new users; Proceedings of eLex*, pp. 286–296.

Vossen, P., I. Maks, R. Segers, H. van der Vliet, M.-F. Moens, K. Hofmann, E. Tjong Kim Sang, and M. De Rijke (2013), Cornetto: a combinatorial lexical semantic database for Dutch, *in* Spyns, P. and J. Odijk, editors, *Essential Speech and Language Technology for Dutch*, Springer, pp. 165–184.

Wauters, L. N, A.E.J.M. Tellings, W.H.J. Van Bon, and A.W. Van Haaften (2003), Mode of acquisition of word meanings: The viability of a theoretical construct, *Applied Psycholinguistics* **24** (03), pp. 385–406, Cambridge University Press.

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

# Appendix A. The 100 words with the highest LL-values in BasiLex-S&V and in BasiLex-SoNaR

| | *BasiLex* | *S&V* | | *BasiLex* | *S&V* | | *BasiLex* | *S&V* |
|---|---|---|---|---|---|---|---|---|
| zijn. to be | 0,02634430 | **0,06393073** | aankijken. to look at | 0,00001701 | **0,00047329** | meneer. mister, sir | 0,00033201 | **0,00088808** |
| hebben. to have | 0,00908403 | **0,02669093** | opschrijven. to write down | 0,00005069 | **0,00061805** | nakijken. to check, examine | 0,00001144 | **0,00019735** |
| zeggen. to say | 0,00490308 | **0,01780124** | opgaan. to go up, ascend, rise | 0,00000579 | **0,00037461** | begrijpen. to comprehend | 0,00021111 | **0,00066414** |
| gaan. to go | 0,00503005 | **0,01668331** | slapen. to sleep | 0,00023792 | **0,00110730** | leeuw. lion | 0,00004454 | **0,00030430** |
| ja. yes | 0,00097914 | **0,00757204** | blijven. to stay, remain | 0,00084453 | **0,00224059** | stuk. broke | 0,00057565 | **0,00125088** |
| meer. lake | 0,00004182 | **0,00263470** | eten. food | 0,00075152 | **0,00201665** | kopen. to buy | 0,00036246 | **0,00091822** |
| he. hey | 0,00000558 | **0,00210705** | gulden. guilder | 0,00000886 | **0,00033503** | aantrekken. to put on, attract | 0,00002995 | **0,00025880** |
| kunnen. can | 0,00502590 | **0,01384594** | jongen. boy | 0,00034845 | **0,00118234** | vallen. to fall | 0,00053819 | **0,00118943** |
| doen. to do | 0,00332906 | **0,01075626** | wakker. awake | 0,00000243 | **0,00025939** | trekken. to draw | 0,00048364 | **0,00110434** |
| kijken. to watch | 0,00221323 | **0,00856412** | heel. complete | **0,00213709** | 0,00025939 | tafel. table | 0,00023792 | **0,00069250** |
| moeten. should, have to | 0,00341256 | **0,01083425** | opstaan. to stand up, rise | 0,00002238 | **0,00387317** | tegenkomen. to meet, encounter | 0,00002717 | **0,00024108** |
| zien. to see | 0,00263138 | **0,00874315** | meenemen. to take along | 0,00008079 | **0,00034743** | paddestoel. fungus | 0,00000307 | **0,00013354** |
| zitten. to sit | 0,00200018 | **0,00741782** | ophouden. to stop | 0,00003617 | **0,00053238** | verdergaan. to go on | 0,00000608 | **0,00015067** |
| komen. to come | 0,00346890 | **0,00957451** | geven. to give | **0,00109146** | 0,00037225 | huis. house | 0,00082315 | **0,00156286** |
| zullen. shall, will | 0,00216468 | **0,00693094** | vliegen. to fly | 0,00027410 | **0,00222700** | drinken. to drink | 0,00016579 | **0,00054597** |
| beetje. (little) bit | 0,00000293 | **0,00115516** | laten. to let, leave | **0,00120713** | 0,00089695 | nemen. to take | 0,00062398 | **0,00127156** |
| mogen. can, may | 0,00118840 | **0,00462890** | klein. small, little | 0,00070748 | **0,00235404** | laat. late | **0,00085797** | 0,00030194 |
| weten. to know | 0,00211135 | **0,00614626** | springen. to jump | 0,00022362 | **0,00078645** | toezien. to look on, supervise | 0,00000193 | **0,00011995** |
| staan. to stand | 0,00245115 | **0,00661660** | kapitein. captain | 0,00004218 | **0,00035984** | opletten. to pay attention | 0,00002767 | **0,00022571** |
| hoor. fine, great | 0,00000193 | **0,00089872** | nadenken. to think, reflect on | 0,00006706 | **0,00043134** | rijden. to ride, drive | 0,00032085 | **0,00079295** |
| denken. to think | 0,00139966 | **0,00450600** | zoeken. to search | 0,00055899 | **0,00134483** | kruipen. to crawl | 0,00009337 | **0,00037225** |
| vinden. to find | 0,00182760 | **0,00527532** | doorgaan. to go on | 0,00003675 | **0,00033325** | opzoeken. to look up | 0,00003396 | **0,00023103** |
| willen. to want, wish | 0,00256039 | **0,00633062** | beer. bear | 0,00006792 | **0,00042129** | meegaan. to go along, accompany | 0,00002195 | **0,00019144** |
| worden. =transitive form | 0,00385116 | **0,00780071** | helpen. to help | 0,00052417 | **0,00126211** | volgen. to follow | **0,00052089** | 0,00013590 |
| nee. no | 0,00084696 | **0,00294373** | opeten. to eat up | 0,00005104 | **0,00036398** | vasthouden. to hold (on) | 0,00002852 | **0,00020858** |
| zijde. side, silk | 0,00001523 | **0,00073564** | zwemmen. to swim | 0,00013190 | **0,00054715** | klimmen. to climb | 0,00010738 | **0,00038525** |
| vragen. to ask | 0,00143104 | **0,00394940** | aankomen. to arrive, gain(weight) | 0,00006742 | **0,00039470** | moed. courage | 0,00002059 | **0,00017963** |
| liggen. to lay | 0,00088506 | **0,00285805** | moeder. mother | 0,00083809 | **0,00169462** | hangen. to hang | 0,00025151 | **0,00063578** |
| roepen. to call | 0,00096370 | **0,00299809** | lijken. to seem, appear, look alike | 0,00051810 | **0,00120715** | uitroepen. to shout | 0,00000572 | **0,00012113** |
| lopen. to walk | 0,00100195 | **0,00304654** | opgave. sum, question, exercise | **0,00037240** | 0,00001773 | brengen. to bring | 0,00033908 | **0,00077345** |
| krijgen. to get | 0,00133267 | **0,00324153** | beet. bite | **0,00041494** | 0,00003368 | terugkomen. to come back | 0,00004404 | **0,00024167** |
| beginnen. to start | 0,00084567 | **0,00240249** | zingen. to sing | 0,00018266 | **0,00062573** | groot. great, large | 0,00153492 | **0,00237531** |
| beter. better | 0,00000965 | **0,00042779** | mooi. beautiful | 0,00059202 | **0,00130169** | omdraaien. to turn around | 0,00002145 | **0,00017903** |
| | | | | | | trol. troll | 0,00000772 | **0,00012822** |

| | *BasiLex* | *SoNaR* | | *BasiLex* | *SoNaR* | | *BasiLex* | *SoNaR* |
|---|---|---|---|---|---|---|---|---|
| groot. great, large | **0,00153492** | 0,00000835 | moeder. mother | **0,00083809** | 0,00022211 | lachen. to laugh | **0,00047828** | 0,00012571 |
| mens. human (being) | **0,00173974** | 0,00017755 | letter. letter (as: a, b, c) | **0,00032257** | 0,00002880 | bladzijde. page | **0,00020025** | 0,00001842 |
| woord. word | **0,00181073** | 0,00024258 | vader. father | **0,00085954** | 0,00024541 | school. school | **0,00070590** | 0,00024929 |
| kijken. to watch, look | **0,00221323** | 0,00043019 | eten. food | **0,00062655** | 0,00013915 | Meneer. Sir | **0,00032679** | 0,00006212 |
| zeggen. to say | **0,00490308** | 0,00182279 | werkwoord. verb | **0,00017380** | 0,00000307 | kleed. carpet, rug | **0,00014649** | 0,00000844 |
| schrijven. to write | **0,00189273** | 0,00035243 | schrift. notebook | **0,00020132** | 0,00000632 | wild. wild | **0,00011095** | 0,00000302 |
| zin. sentence | **0,00121099** | 0,00014602 | kind. child | **0,00146736** | 0,00061338 | zetten. to put, place | **0,00087284** | 0,00037117 |
| lezen. to read | **0,00125381** | 0,00018367 | weten. to know | **0,00211135** | 0,00104781 | hond. dog | **0,00031921** | 0,00006647 |
| roepen. to call | **0,00096370** | 0,00013658 | heel. complete | **0,00213709** | 0,00108860 | ver. far | **0,00060081** | 0,00021118 |
| beet. bite | **0,00041193** | 0,00001128 | knikken. to nodd | **0,00024042** | 0,00001694 | kleur. colour | **0,00036096** | 0,00008873 |
| heten. to be called | **0,00037976** | 0,00000821 | opdracht. assignment | **0,00048929** | 0,00009978 | vertellen. to tell | **0,00080277** | 0,00033723 |
| gaan. to go | **0,00503005** | 0,00258296 | tekening. drawing | **0,00028554** | 0,00002907 | inwonen. to inhabit | **0,00011489** | 0,00000499 |
| juf. miss* | **0,00036353** | 0,00000723 | water. water | **0,00069926** | 0,00019988 | euro. euro | 0,00026287 | **0,00076488** |
| opgave. sum, question, exercise | **0,00037240** | 0,00000846 | staan. to stand | **0,00245115** | 0,00135508 | hok. pen (animals) | **0,00011389** | 0,00000511 |
| oma. gran(ny) | **0,00037933** | 0,00001430 | meisje. girl | **0,00044310** | 0,00008686 | loop. course | **0,00010666** | 0,00000401 |
| tekst. text | **0,00073665** | 0,00010442 | plaat. plate | **0,00032100** | 0,00004445 | nieuw. new | **0,00010759** | 0,00000415 |
| horen. to hear | **0,00138293** | 0,00039289 | procent. percent | 0,00003796 | **0,00046879** | boer. farmer | **0,00024900** | 0,00004520 |
| opa. gramps | **0,00032207** | 0,00000956 | antwoord. answer | **0,00048285** | 0,00010775 | seizoen. season | 0,00003081 | **0,00030617** |
| bedenken. to think about | **0,00052832** | 0,00005144 | kaart. map | **0,00047141** | 0,00010324 | papa. daddy | **0,00021740** | 0,00003487 |
| meester. master* | **0,00048850** | 0,00004269 | maken. to make | **0,00286901** | 0,00170168 | lopen. to walk | **0,00100195** | 0,00048880 |
| ding. thing | **0,00051181** | 0,00005010 | goed. good, well | **0,00330840** | 0,00204922 | bedrijf. company | 0,00009351 | **0,00043828** |
| zien. to see | **0,00262895** | 0,00118138 | nee. no | **0,00084410** | 0,00030717 | tante. aunt | **0,00017172** | 0,00002084 |
| laat. late | **0,00084910** | 0,00017559 | denken. to think | **0,00139966** | 0,00065895 | zullen. shall, will | 0,00216468 | **0,00258144** |
| jong. young | **0,00035081** | 0,00001871 | mam. mom | **0,00014856** | 0,00000531 | komen. to come | **0,00346890** | 0,00241970 |
| klas. class | **0,00039477** | 0,00002951 | pap. dad | **0,00016321** | 0,00000801 | pakken. to fetch, take | **0,00051316** | 0,00018281 |
| doen. to do | **0,00332906** | 0,00169637 | boos. angry | **0,00028475** | 0,00003956 | snappen. to comprehend | **0,00018130** | 0,00002469 |
| vragen. to ask | **0,00143104** | 0,00048435 | mama. mummy (as: mum) | **0,00028289** | 0,00003980 | huis. house | **0,00082315** | 0,00037670 |
| worden. =transitive form | 0,00385116 | **0,00635424** | lekker. delicious | **0,00036260** | 0,00006906 | schrikken. to be frightened | **0,00020332** | 0,00003276 |
| jaar. year | 0,00108853 | **0,00259088** | letten. to pay attention | **0,00019567** | 0,00001538 | kruis. cross | **0,00015571** | 0,00001732 |
| leuk. nice, cute | **0,00065114** | 0,00012647 | fluisteren. to whisper | **0,00016078** | 0,00000819 | zaak. business | 0,00006127 | **0,00035960** |
| rennen. to run | **0,00031513** | 0,00002208 | hopen. to hope | **0,00016207** | 0,00000901 | juffrouw. miss | **0,00011224** | 0,00000708 |
| dier. animal | **0,00060045** | 0,00011212 | vinden. to find | **0,00182631** | 0,00098290 | stom. stupid | **0,00015528** | 0,00001881 |
| zitten. to sit | **0,00199832** | 0,00090345 | vullen. to fill | **0,00033308** | 0,00006051 | lucht. air | **0,00030048** | 0,00007694 |
| | | | | | | raar. weird | **0,00015871** | 0,00001999 |