Testing the Robustness of Laws of Polysemy and Brevity versus Frequency

Antoni Hernández-Fernández[2], Bernardino Casas[1], Ramon Ferrer-i-Cancho[1], and Jaume Baixeries[1]

[1] Complexity & Quantitative Linguistics Lab, Laboratory for Relational Algorithmics, Complexity and Learning (LARCA), Departament de Ciències de la Computació, Universitat Politècnica de Catalunya, Barcelona, Catalonia.
{*bcasas*, *jbaixer*, *rferrericancho*}@cs.upc.edu
[2] Complexity & Quantitative Linguistics Lab, Laboratory for Relational Algorithmics, Complexity and Learning (LARCA), Institut de Ciències de l'Educació, Universitat Politècnica de Catalunya, Barcelona, Catalonia.
antonio.hernandez@upc.edu

Corresponding author: antonio.hernandez@upc.edu

Abstract. The pioneering research of G. K. Zipf on the relationship between word frequency and other word features led to the formulation of various linguistic laws. Here we focus on a couple of them: the meaning-frequency law, i.e. the tendency of more frequent words to be more polysemous, and the law of abbreviation, i.e. the tendency of more frequent words to be shorter. Here we evaluate the robustness of these laws in contexts where they have not been explored yet to our knowledge. The recovery of the laws again in new conditions provides support for the hypothesis that they originate from abstract mechanisms.
Keywords: Zipf law, polysemy, brevity, word frequency.

1 Introduction

The linguist George Kingsley Zipf (1902-1950) is known for his investigations on statistical laws of language [20, 21]. Perhaps the most popular one is **Zipfs law for word frequencies** [20], that states that the frequency of the i-th most frequent word in a text follows approximately

$$f \propto i^{-\alpha} \tag{1}$$

where f is the frequency of that word, i their rank or order and α is a constant ($\alpha \approx 1$). Zipfs law for word frequencies can be explained by information theoretic models of communication and is a robust pattern of language that presents invariance with text length [9] but dependency with respect to the linguistic units considered [5]. The focus of the current paper are a couple of linguistic laws that are perhaps less popular:

– **Meaning-frequency law** [19], the tendency of more frequent words to be more polysemous

– **Zipfs law of abbreviation** [20], the tendency of more frequent words to be shorter or smaller.

These laws are examples of laws that where the predictor is word frequency and the response is another word feature. These laws are regarded as universal although the only evidence of their universality is that they hold in every language or condition where they have been tested. Because of their generality, these laws have triggered modelling efforts that attempt to explain their origin and support their presumable universality with the help of abstract mechanisms or linguistic principles, e.g., [8]. Therefore, investigating the conditions under which these laws hold is crucial.

In this paper we contribute to the exploration of different definitions of word frequency and word polysemy to test the robustness of these linguistic laws in English (taking into account in our analysis only content words (nouns, verbs, adjectives and adverbs)). Concerning word frequency, in this preliminary study, we consider three major sources of estimation: the CELEX lexical database [3], the CHILDES database [16] and the SemCor corpus [1]. The estimates from the CHILDES database are divided into four types depending on the kind of speakers: children, mothers, fathers and investigators. Concerning polysemy, we consider two related measures: the number of synsets of a word according to WordNet [6], that we refer to as WordNet polysemy, and the number of synsets of WordNet that have appeared in the SemCor corpus, that we refer to as SemCor polysemy. These two measures of polysemy allow one to capture two extremes: the full potential number of synsets of a word (WordNet polysemy) and the actual number of synsets that are used (SemCor polysemy), being the latter a more conservative measure of word polysemy motivated by the fact that, in many cases, the number of synsets of a word overestimates the number of synsets that are known to an average speaker of English. In this study, we assume the polysemy measure provided by Wordnet, although we are aware of the inherent difficulties of borrowing this conceptual framework (see [12, 15]). Concerning word length we simply consider orthographic length. Therefore, the SemCor corpus contains SemCor polysemy and SemCor frequency, as well as the length of its lemmas, and the CHILDES database contains CHILDES frequency, the length of its lemmas, and has been enriched with CELEX frequency, WordNet polysemy, and SemCor polysemy. The conditions above lead to $1 + 2 \times 2 = 5$ major ways of investigating the meaning-frequency law and to $1 + 2 = 3$ ways of investigating the law of abbreviation (see details in Section 3). The choice made in this preliminary study should not be considered a limitation, since we plan to extend the range of data sources and measures in future studies (we explain these possibilities in Section 5).

In this paper, we investigate these laws qualitatively using measures of correlation between two variables. Thus, the law of abbreviation is defined as a significant negative correlation between the frequency of a word and its length. The meaning-frequency law is defined as a significant positive correlation between

[1] `http://multisemcor.fbk.eu/semcor.php`

the frequency of a word and its number of synsets, a proxy for the number of meanings of a word. We adopt these correlational definitions to remain agnostic about the actual functional dependency between the variable, which is currently under revision for various statistical laws of language [1]. We will show that a significant correlation of the right sign is found in all the combinations of conditions mentioned above, providing support for the hypothesis that these laws originate from abstract mechanisms.

2 Materials

In this section we describe the different corpora and tools that have been used in this paper. We first describe the WordNet database and CELEX corpus, which have been used to compute polysemy and frequency measures. Then, we describe the two different corpora that are analyzed in this paper: SemCor and CHILDES.

2.1 Lexical database WordNet

The WordNet database [6] can be seen as a set of senses (also called synset) and relationships among them, where a synset is the representation of an abstract meaning and is defined as a set of words having (at least) the meaning that the synset stands for. Apart from this pair of sets, a relationship between both is also contained. Each pair word-synset is also related to a syntactical category. For instance, the pair *book* and the synset *a written work or composition that has been published* are related to the category *noun*, whereas the pair *book* and synset *to arrange for and reserve (something for someone else) in advance* are related to the category *verb*. WordNet has 155,287 lemmas and 117,659 synsets and contains only four main syntactic categories: nouns, verbs, adjectives and adverbs.

2.2 CELEX corpus

CELEX [3] is a text corpora in Dutch, English and German, but in this paper we only use the information in English. For each language, CELEX contains detailed information on orthography, phonology, morphology, syntax (word class) and word frequency, based on resent and representative text corpora.

2.3 SemCor corpus

SemCor is a corpus created at Princeton University composed of 352 texts which are a subset of the English Brown Corpus. All words in the corpus have been sintactically tagged using Brill's part of speech tagger. The semantical tagging has been done manually, mapping all nouns, verbs, adjectives and adverbs, to their corresponding synsets in the WordNet database.

SemCor contains $676,546$ tokens, $234,136$ of which are tagged. In this article we only analyze content words (nouns, verbs, adjectives and adverbs), thus it yields $23,341$ different tagged lemmas that represent only content words.

We use the SemCor corpus to obtain a new measure of polysemy.

SemCor corpus is freely available for download at `http://web.eecs.umich.edu/~mihalcea/downloads.html\#semcor` (accessed 22 april 2016).

2.4 CHILDES database

The CHILDES database [16] is a set of corpora of transcripts of conversations between children and adults. The corpora included in this database are in different languages, and contains conversations when the children were between 12 and 65 months old, approximately. In this paper we have studied the conversations of 60 children in English (detailed information on these conversations can be found in [4]).

We analyze syntactically every conversation of the selected corpora of CHILDES using Treetagger in order to obtain the lemma and part-of-speech for every word. We have for each word from CHILDES said for each role: lemma, part-of-speech, frequency (number of times that this word is said by this role), number of synsets (according to both SemCor or WordNet), and the word length. We only have taken into account content words (nouns, verbs, adjectives and adverbs). Figure 1 shows the amount of different lemmas obtained from the selected corpora of CHILDES and the amount of analyzed lemmas in this paper for each category. The amount of analyzed lemmas from this corpus is smaller than the total number of lemmas because we have only analyzed those lemmas that are also present in the SemCor corpus.

Role	Tokens	# Lemmas	# Analyzed Lemmas
Child	1,358,219	7,835	4,675
Mother	2,269,801	11,583	6,962
Father	313,593	6,135	4,203
Investigator	182,402	3,659	2,775

Fig. 1. Number of tokens, lemmas and analyzed lemmas obtained from CHILDES conversations for each role.

3 Methods

In this paper we compute the relationship between three variables that are related to every lemma: length, frequency and polysemy.

3.1 Length

For the length, we compute the number of letters of the lexical item. Blanks, separation characters and the like have not been taken into consideration.

3.2 Frequency

We have calculated the frequency from three different sources:

- **SemCor frequency**. We use the frequency of each pair *lemma, syntactic category* that is present in the SemCor dataset.
- **CELEX frequency**. We use the frequency of each pair *lemma, syntactic category* that is present in the CELEX lexicon.
- **CHILDES frequency**. For each pair *lemma, syntactic category* that appears in the CHILDES database, we compute its frequency according to each role: child, mother, father, investigator. For example, for the pair *book, noun* we count four different frequencies: the number of times that this pair appears uttered by a child, a mother, a father and an investigator, respectively.

SemCor frequency can only be analyzed in the SemCor corpus, whereas CELEX and CHILDES frequencies are only analyzed in the CHILDES corpora.

3.3 Polysemy

We have calculated the polysemy from two different sources:

- **SemCor polysemy**. For each pair *lemma, syntactic category* we compute the number of different synsets with which this pair has been tagged in the SemCor corpus. This measure is analyzed in the SemCor corpus and in the CHILDES corpus.
- **WordNet polysemy**. For each pair *lemma, syntactic category* we consider the number of synsets according to the WordNet database. This measure is only analyzed in the CHILDES corpus.

We are aware that using a SemCor polysemy measure in the CHILDES corpus or using Wordnet polysemy in both SemCor and CHILDES corpora induces a bias. In the former case, because we are assuming that the same meanings that are used in written text are also used in spoken language. In the latter case, because we are using all possible meanings of a word. An alternative would have been to tag manually all corpora (which is currently an unavailable option) or use an automatic tagger. But also in this case, the possibility of biases or errors would be present. We have performed these combinations for the sake of completeness, and also assuming their limitations.

3.4 Statistical Methods

To compute the relationship between (1) frequency and polysemy and (2) frequency and length. Since frequency and polysemy have more than one source, we have computed all available combinations. In this paper, for the SemCor corpus we analyze the relationship between:

1. SemCor frequency and SemCor polysemy.

2. SemCor frequency and lemma length in the SemCor corpus.

As for the CHILDES corpora, the availability of different sources for frequency and polysemy yields the following combinations:

1. CELEX frequency and SemCor polysemy.
2. CELEX frequency and WordNet polysemy.
3. CHILDES frequency and SemCor polysemy.
4. CHILDES frequency and WordNet polysemy.
5. CHILDES frequency and lemma length in the CHILDES corpus.
6. CELEX frequency and lemma length in the CHILDES corpus.

For each combination of two variables, we compute:

1. **Correlation test**. Pearson, Spearman and Kendall correlation tests, using the `cor.test` standardized R function.
2. **Plot**, in logarithmic scale, that also shows the density of points.
3. **Nonparametric regression**, using the `locpoly` standarized R function, which has been overlapped in the previous plot.

We remark that the analysis for the CHILDES corpora has been segmented by role.

4 Results

We analyze the relationship between (1) frequency and polysemy and (2) frequency and length separately in two different corpora (SemCor and CHILDES).

In both corpora, we have computed a correlation test and a nonparametric regression, which has been plotted alongside with the values of the two variables that are analyzed.

For the SemCor corpus, we have analyzed the relationship between the SemCor frequency and the SemCor polysemy and the relationship between the SemCor frequency and the length of lemmata.

As for the CHILDES corpora, we have analyzed the relationship between two different measures of frequency (CHILDES and CELEX) versus two different measures of polysemy (WordNet and SemCor) and also, the relationship between two different measures of frequency (CHILDES and CELEX) and the length of lemmas. The analysis of individual roles (child, mother, father and investigator) does not show any significant difference between them. In **all** cases we have that:

1. The value of the correlation is *positive* for the relationships frequency-polysemy (see Figure 2), and *negative* for the relationships frequency-length (see Figure 4) for all types of correlation: Pearson, Spearman and Kendall. We remark that the p-value is *near zero* in all cases. This is, all correlations are significant.

2. The nonparametric regression function draws a line with a *positive* slope for the frequency-polysemy relationship (see Figure 3), and *negative* slope for the frequency-length relationship (see Figure 5). When we say that *it draws a line*, we mean that this function is a quasi-line in the central area of the graph, where most of the points are located. This tendency is not maintained at the extreme parts of graph, where the density of points is significantly lower.

Corpus	ρ	ρ_S	τ_K	**Corpus length**
SemCor frequency versus SemCor polysemy				
SemCor	0.209	0.627	0.555	23341
CHILDES frequency versus CELEX polysemy				
CHILDES (children)	0.084	0.249	0.177	4675
CHILDES (mothers)	0.081	0.281	0.202	6962
CHILDES (fathers)	0.084	0.279	0.202	4203
CHILDES (investigators)	0.062	0.211	0.153	2775
CELEX frequency versus WordNet polysemy				
CHILDES (children)	0.073	0.353	0.249	4406
CHILDES (mothers)	0.085	0.366	0.261	6577
CHILDES (fathers)	0.089	0.373	0.264	3989
CHILDES (investigators)	0.075	0.341	0.24	2654
CHILDES frequency versus SemCor polysemy				
CHILDES (children)	0.211	0.230	0.178	4675
CHILDES (mothers)	0.186	0.252	0.197	6962
CHILDES (fathers)	0.201	0.256	0.200	4203
CHILDES (investigators)	0.189	0.219	0.171	2775
CELEX frequency versus SemCor polysemy				
CHILDES (children)	0.201	0.607	0.477	4406
CHILDES (mothers)	0.197	0.602	0.474	6577
CHILDES (fathers)	0.226	0.595	0.463	3989
CHILDES (investigators)	0.228	0.585	0.451	2654

Fig. 2. Summary of the analysis of the correlation between the frequency and polysemy of each lemma. Three statistics are considered: the sample Pearson correlation coefficient (ρ), the sample Spearman correlation coefficient (ρ_S) and the sample Kendall correlation tau (τ_K). All correlation tests indicates a significant negative correlation with p-values under 10^{16}

5 Discussion and Future Work

In this paper, we have reviewed two linguistic laws that we owe to Zipf's ([19], [20]) and that have probably been shadowed by the best-known Zipf's law for

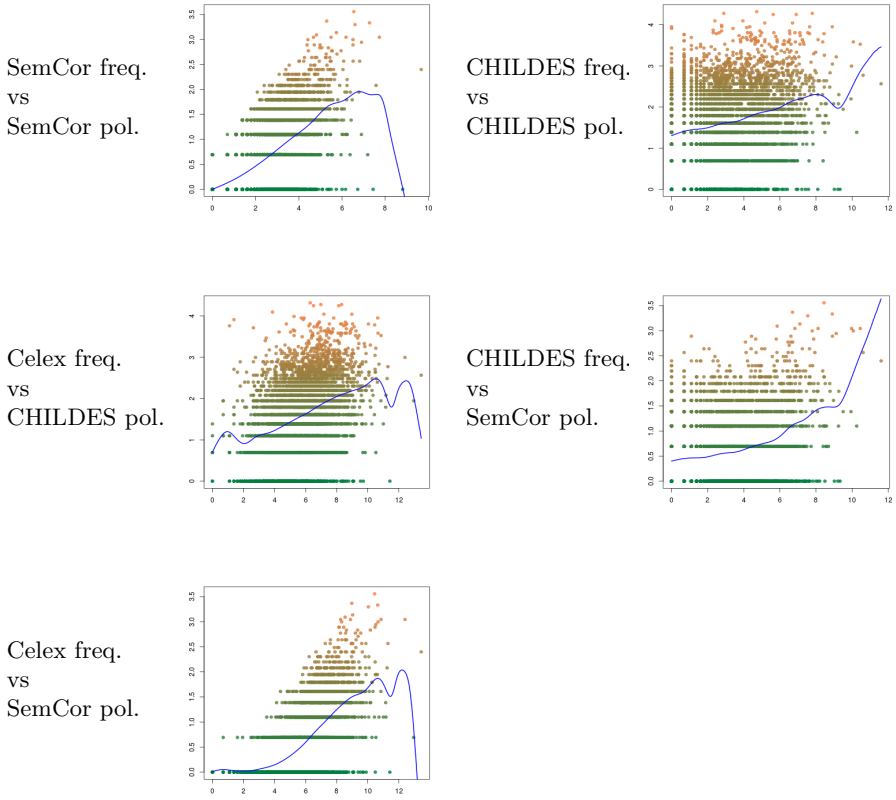

Fig. 3. Graphics of the relation between frequency (x-axis) and polysemy (y-axis), both in logarithmic scale. The color indicates the density of points: dark green is the highest possible density. The blue line is the nonparametric regression performed over the logarithmic values of frequency and polysemy. In the case of graphs concerning the CHILDES corpus, we show the graphs only for children

word frequencies ([20]). Our analysis of the correlation between brevity (measured in number of characters) and polysemy (number of synsets) versus lemma frequency was conducted with three tests with varying assumptions and robustness. Pearson's method supposes input vectors approximately normally distributed while Spearman's is a non-parametric test that does require vectors being approximately normally distributed [2]. Kendall's tau is more robust to extreme observations and to non-linearity compared with the standard Pearson product-moment correlation [17]. Our analysis confirm that a positive correlation between the frequency of the lemmas and the number of synsets (consistent with the meaning-frequency law) and a negative correlation between the length of the lemmas and their frequency (consistent with the law of abbreviation) arises under different definitions of the variables. Interestingly, we have not found any

Corpus	ρ	ρ_S	τ_K	Corpus length
SemCor frequency versus lemma length				
SemCor	−0.062	−0.301	−0.229	23341
CHILDES frequency versus lemma length				
CHILDES (children)	−0.099	−0.324	−0.24	4675
CHILDES (mothers)	−0.076	−0.373	−0.278	6962
CHILDES (fathers)	−0.092	−0.366	−0.277	4203
CHILDES (investigators)	−0.096	−0.318	−0.242	2775
CELEX frequency versus lemma length				
CHILDES (children)	−0.091	−0.132	−0.095	4406
CHILDES (mothers)	−0.084	−0.124	−0.089	6577
CHILDES (fathers)	−0.087	−0.142	−0.102	3989
CHILDES (investigators)	−0.099	−0.172	−0.126	2654

Fig. 4. Summary of the analysis of the correlation between the frequency and the lemma length. Three statistics are considered: the sample Pearson correlation coefficient (ρ), the sample Spearman correlation coefficient (ρ_S) and the sample Kendall correlation tau (τ_K). All correlation tests indicates a significant negative correlation with p-values under 10^{16}

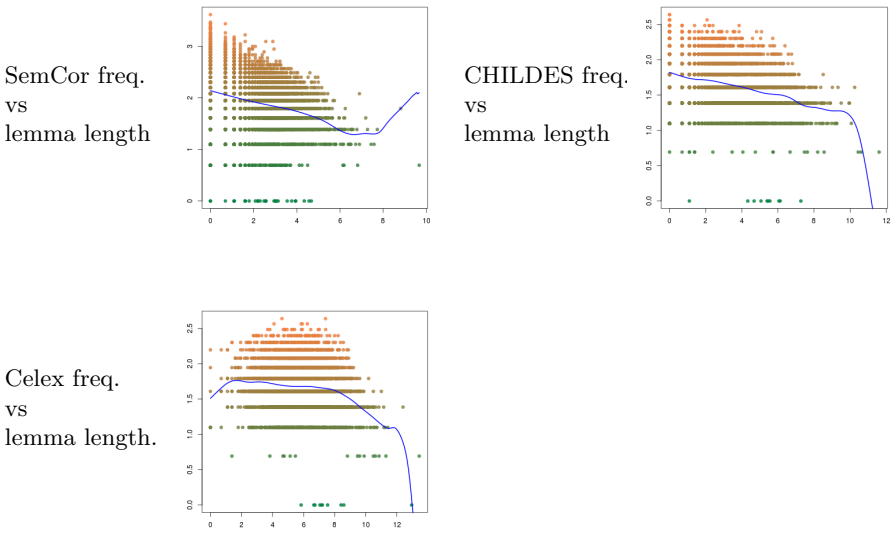

Fig. 5. Graphics of the relation between frequency (x-axis) and lemma length (y-axis), both in logarithmic scale. The color indicates the density of points: dark green is the highest possible density. The blue line is the nonparametric regression performed over the logarithmic values of frequency and lemma length. In the case of graphs concerning the CHILDES corpus, we show the graphs only for children

remarkable qualitative difference in the analysis of correlations for the different speakers (roles) in the the Childes database, suggesting that both child speech and the child-directed-speech (the so-called *motherese*) seem to show the same general statistical biases in the use of more frequent words (that tend to be shorter and more polysemous). With this regard, our results agree with Zipf's pioneering discoveries, independently from the corpora analyzed and independently from the source used to measure the linguistic variables.

Our work offers many possibilities for future research.

First, the analysis of more extensive databases, e.g., Wikipedia in the case of word-length versus frequency. Second, the use of more fine-grained statistical techniques that allow: (1) to unveil differences between sources or between kinds of speakers, (2) to verify that the tendencies that are shown in this preliminary study are correct, and (3) to explain the variations that are displayed in the graphics and to characterize the words that are in the part of the graphics in which our hypotheses hold. Third, considering different definitions of the same variables. For instance, a limitation of our study is the fact that we define word length using graphemes. An accurate measurement of brevity would require detailed acoustical information that is missing in raw written transcripts[10] or using more sophisticated methods of computation, for instance, to calculate number of phonemes and syllables according to [1]. However, the relationship between the duration of phonemes and graphemes is well-known and in general longer words has longer durations: grapheme-to-phoneme conversion is still a hot topic of research, due to the ambiguity of graphemes with respect to their pronunciation that today supposes a difficulty in speech technologies [18]. In order to improve the frequency measure, we would consider the use of alternative databases, e.g., the frequency of English words in Wikipedia [11]. Forth, our work can be extended including other linguistic variables such as homophony, i.e. words with different origin (and *a priori* different meaning) that have converged to the same phonological form. Actually, Jespersen (1933) suggested a connection between brevity of words and homophony [13], confirmed by Ke(2006) more recently [14] and reviewed by Fenk-Oczlon and Fenk (2010) that outline the "*strong association between shortness of words, token frequency and homophony*"[7]. In fact, the study of different types of polysemy and its multifaceted implications in linguistic networks is descent as future work, as well as the direct study of human voice, because every linguistic phenomenon or candidate for a language law, could be camouflaged or diluted in our transcripts of oral corpus by writing technology, a technology that has been very useful during the last five thousand years, but that prevents us from being close to the acoustic phenomenon of language ([10]).

Acknowledgments

The authors thank Pedro Delicado and the reviewers for their helpful comments. This research work has been supported by the SGR2014-890 (MACDA) project of the Generalitat de Catalunya, and MINECO project APCOM (TIN2014-57226-P).

Bibliography

[1] Altmann, E.G., Gerlach, M.: Statistical Laws in Linguistics, pp. 7–26. Springer International Publishing, Cham (2016), http://dx.doi.org/10.1007/978-3-319-24403-7_2

[2] Baayen, R.H.: Analyzing Linguistic Data: A practical introduction to statistics using R. Cambridge University Press, Cambridge (2007)

[3] Baayen, R.H., Piepenbrock, R., Gulikers, L.: CELEX (1996)

[4] Baixeries, J., Elvevåg, B., Ferrer-i-Cancho, R.: The evolution of the exponent of zipf's law in language ontogeny. PLoS ONE 8(3) (2013)

[5] Corral, A., Boleda, G., Ferrer-i Cancho, R.: Zipf's law for word frequencies: Word forms versus lemmas in long texts. PLoS ONE 10(7), 1–23 (07 2015)

[6] Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Pres, Cambridge, MA (1998)

[7] Fenk-Oczlon, G., Fenk, A.: Frequency effects on the emergence of polysemy and homophony. International Journal Information Technologies and Knowledge 4(2), 103–109 (2010)

[8] Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M.J., Semple, S.: Compression as a universal principle of animal behavior. Cognitive Science 37(8), 15651578 (2013)

[9] Font-Clos, F., Boleda, G., Corral, A.: A scaling law beyond zipf's law and its relation to heaps' law. New Journal of Physics 15(9), 093033 (2013), http://stacks.iop.org/1367-2630/15/i=9/a=093033

[10] Gonzalez Torre, I., Luque, B., Lacasa, L., Luque, J., Hernandez-Fernandez, A.: Emergence of linguistic laws in human voice. In preparation (2016)

[11] Grefenstette, G.: Extracting weighted language lexicons from wikipedia. In: Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)

[12] Ide, N., Wilks, Y.: Making Sense About Sense, pp. 47–73. Springer Netherlands, Dordrecht (2006), http://dx.doi.org/10.1007/978-1-4020-4809-8_3

[13] Jespersen, O.: Monosyllabism in english. In: Linguistica: Selected Writings of Otto Jespersen. pp. 574–598. George Allen and Unwin LTD, London, UK (2007)

[14] Ke, J.: A cross-linguistic quantitative study of homophony. Journal of Quantitative Linguistics pp. 129–159 (2006)

[15] Kilgarriff, A.: Dictionary word sense distinctions: An enquiry into their nature. Computers and the Humanities 26(5), 365–387 (1992), http://dx.doi.org/10.1007/BF00136981

[16] MacWhinney, B.: The CHILDES project: tools for analyzing talk, vol. 2: the database. Lawrence Erlbaum Associates, Mahwah, NJ, 3rd edn. (2000)

[17] Newson, R.: Parameters behind nonparametric statistics: Kendalls tau, somers d and median differences. Stata Journal 2(1), 45–64 (2002)
[18] Razavi, M., Rasipuram, R., Magimai.-Doss, M.: Acoustic data-driven grapheme-to-phoneme conversion in the probabilistic lexical modeling framework. Speech Communication 80 (2016)
[19] Zipf, G.K.: The Meaning-Frequency Relationship of Words. Journal of General Psychology 1945(33), 251–256 (1945)
[20] Zipf, G.K.: Human behaviour and the principle of least effort. Addison-Wesley, Cambridge (MA), USA (1949)
[21] Zipf, G.K.: The Psycho-Biology of Language: an Introduction to Dynamic Psychology. MIT Press, Cambridge, MA, USA (1968), originally published in 1935 by Houghton Mifflin - Boston - MA - USA