Comparing syllable frequencies in corpora of written and spoken language

Barbara Samlowski [1], *Bernd Möbius* [2], *Petra Wagner* [3]

[1] Division of Language and Speech Communication, University of Bonn, Germany
[2] Department of Computational Linguistics and Phonetics, Saarland University, Germany
[3] Faculty of Linguistics and Literary Studies, Bielefeld University, Germany
bsa@sk.uni-bonn.de, moebius@coli.uni-saarland.de, petra.wagner@uni-bielefeld.de

Abstract

In this study, various German language corpora were compared in order to discover the extent to which syllable frequencies remain stable across different contexts and modalities. Although considerable differences in relative frequency were found among the more common syllables, rank numbers proved to be more robust. Variation across corpora was mostly due to vocabulary characteristics of particular corpus domains rather than to systematic differences between spoken and written language. The results indicate that syllable frequencies in written corpora can be taken as a rough estimate for their frequency in spoken language.

Index Terms: syllabary, syllable frequencies, spoken and written language corpora

1. Introduction

Estimating syllable frequencies in a language on the basis of speech corpora is an important task for many areas of linguistics, phonetics, and speech technology. The mental syllabary theory in the field of psycholinguistics, for instance, postulates that the human brain stores and accesses whole articulatory routines for frequent syllables, whereas rare or unknown syllables have to be assembled segment by segment [1]. Here, groups of frequent and rare syllables need to be determined in order to analyze whether there are noticeable syllable frequency effects on phenomena such as production latencies or degrees of coarticulation (e.g. [2, 3]). In languages such as English and German, however, syllable frequencies have a highly uneven distribution, and spoken language corpora tend to be too small to adequately represent the rarer syllable types [4, 5]. An alternative approach is to analyze syllable frequencies on the basis of automatically transcribed corpora of written language, which are easier to assemble and are therefore able to provide large amounts of data.

Because spoken and written language are used in different circumstances and often require different degrees of formality, they tend to vary in discourse structure, grammatical constructions, and vocabulary [6, 7, 8, 9]. Of these three areas, only vocabulary differences have a direct influence on syllable frequencies. In the present study, syllable frequencies from various German corpora of spoken and written language were obtained and analyzed in order to determine which syllables were strongly over- or underrepresented in individual corpora, and which remained stable across different contexts and modalities.

2. Methods

2.1. Corpora analyzed

Syllable frequency lists were created from five German language corpora: two corpora of written and three of spoken language. The largest database analyzed for the present study is the DEWAC corpus, a collection of texts crawled from the internet containing more than 1.5 billion running word forms [10]. The other written language corpus, referred to here as the Leipzig corpus, comprises approximately 170 million words from newspaper articles [11].

All spoken-language corpora were analyzed on the basis of orthographic representations. The Europarl corpus consists of transcriptions from European Parliament proceedings [12]. With a total of nearly 40 million words, it is the largest spoken language corpus investigated here, although the style of speech is comparatively formal and planned. The other two spoken language corpora have a more spontaneous speaking style. One, referred to here as GF ("Gespräche im Fernsehen", around 450 000 words), is made up of transcripts from TV talk shows and discussion programs [13], while the other ("Verbmobil", around 300 000 words) contains transcripts from simulated appointment-making dialogues [14].

2.2. Retrieval of syllable frequencies

Extracting syllable frequencies from the various corpora proved to be a long and complex process. In a first step, annotations added to the actual corpus texts had to be removed, and character encoding formats needed to be unified. In this case, the encoding ISO-8859-1 is required by the programs used for the further normalization and transcription process. A number of characters from encoding schemes such as UTF-8, which is used by the Europarl corpus to correctly depict diacritics in foreign names and terms, cannot be automatically transformed as they have no direct counterparts. As a consequence, before the encoding format could be converted, a script had to be written to replace the letters in question with their closest non-diacritic equivalents.

The two written language corpora as well as the Europarl corpus contain elements such as numbers, abbreviations, and special characters that have various possible pronunciations depending on their meaning in the sentence. These elements had to be disambiguated before phonological transcriptions could be generated. As such a text normalization process is an important part of text-to-speech synthesis systems [15], ready-made tools exist for this task. Here, the German preprocessing module of the Festival speech synthesis program was used [16]. While not being free from error and having a comparatively low processing speed, this program includes a finely differentiated set of rules and, for the most part, delivers accurate results. In cases where unexpected contexts led to internal program errors, the sentence in question was left unprocessed. Additional scripts were written to reduce some of the most common mistakes in the Festival preprocessing. However, several sources of error had to remain uncorrected.

The automatic transcriptions were made with the program txt2pho from the speech synthesis system Hadifix [17]. In addition to a pronunciation dictionary of over 80 000 entries,

txt2pho contains rules for analyzing inflected forms and compounds as well as a set of context-sensitive rewrite rules to transcribe unknown words. Its output consists of a phonetic transcription in the form of SAMPA characters [18].

The transcribed words were split into syllables with the help of a statistical tagger based on joint n-gram models [19]. The training lexicon for this tagger was compiled from the Festival Bomp dictionary, which contains nearly 150 000 entries [20].

2.3. Corpus Comparison

Comparing frequency information from different corpora is not always a straightforward task. Uneven syllable frequency distributions can lead to consequences which are similar to those already shown for word frequencies (e.g. [21, 22]). For instance, the number of different types as well as the mean type frequency tend to increase with corpus size. Also, diagrams depicting frequencies and their distributions can easily be dominated by the high frequency of the common syllables or the large amount of rare ones, especially if corpora of different sizes are compared in a single diagram. Relative numbers, logarithmic scaling, and focusing on particular frequency sections can reduce these effects, at the expense of emphasizing some aspects and distorting or ignoring others. In this study, relative syllable frequencies, frequency rank differences, and relative frequency ratios were examined to determine in how far written corpora can be used to accurately estimate syllable frequencies of spoken language. As a basis of comparison, mean relative frequency and rank values were calculated for each syllable. Because the ranks and frequencies were not weighted according to corpus size, each database had an equal influence on the computed reference values. All diagrams included here were created using the statistics program R and its module zipfR [23, 24].

3. Results

3.1. Relative frequencies

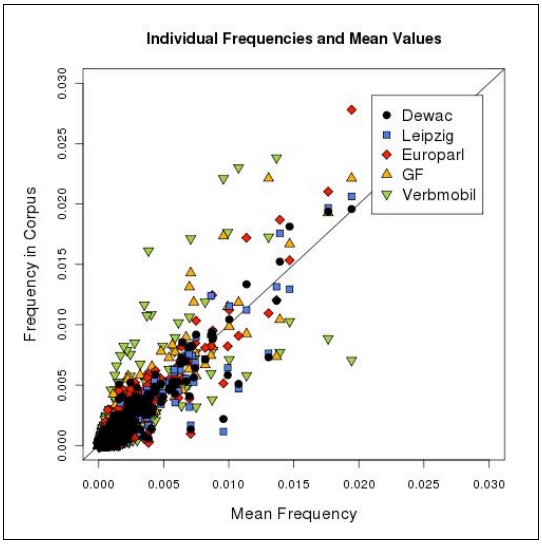

Figure 1: Scatter plot of relative syllable frequencies in five corpora against the mean of these values

Figure 1 directly contrasts the individual syllable frequencies. For each syllable, the vertical axis shows the relative frequencies in the various corpora, while the horizontal axis represents the mean of these five values.

While a dense cluster of rare syllables is formed in the lower left-hand corner of the diagram, considerable differences become visible among the higher frequencies. Especially the Verbmobil corpus contains a number of underrepresented as well as several overrepresented syllables. The other conversational corpus GF shows similar tendencies, but to a lesser extent. For the most part, the frequencies from the written language corpora as well as the Europarl corpus remain closer to their mean values, with only a few instances of highly over- or underrepresented syllables.

3.2. Rank differences

Despite large differences in actual relative frequency among the very frequent syllables, their rank numbers remain rather similar. Examining the syllables which occur among the 100 most frequent in at least one of the databases, 52 are shared by all five corpora, 70 by all but the Verbmobil corpus, and 91 by the two written language corpora Dewac and Leipzig. For all syllables appearing among the top 100 in at least one of the databases, rank differences were computed in order to determine how robust their placing is across various corpora. The two syllables [p E:] and [I] did not appear at all in the Verbmobil corpus and were therefore not included in the analysis, leaving 169 syllables in total. For these syllables, mean ranks were calculated, from which rank numbers in the individual corpus were then subtracted. Only five syllables proved to be more than 1000 ranks over or under the mean in one or more of the corpora. The others are plotted in Figure 2, with the syllables along the horizontal axis and their difference from the mean on the vertical axis.

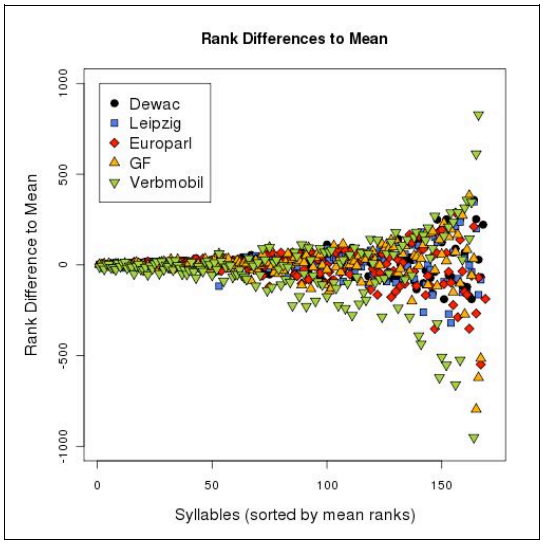

Figure 2: Dot chart with differences of syllable ranks from five corpora to their mean values

For many of the more common syllables, rank numbers remain close together. Verbmobil once again shows the greatest difference from the mean ranks. Of the syllables analyzed, 158 are less than 500 ranks above or below the mean in all databases, 96 have a rank difference of less than 100, and 63 of less than 50. Among the most frequent syllables there is the least divergence in rank, but differences become larger as the mean rank numbers increase.

3.3. Relative frequency ratios

Kilgarriff proposes a way of finding particular key words in one database compared with another, a method which is independent of corpus size and with which it is possible to focus on various frequency ranges [25]. For this purpose, ratios of relative word frequencies are computed. In order to also be able to analyze words that appear in only one corpus without having to divide by zero, a constant is first added to all frequencies. Depending on the value added, the calculated ratios focus on different levels of frequency. Adding 1000 to the frequency counts per million words puts the emphasis on the common types, while a smaller constant such as 100 or 1 emphasizes less frequent words.

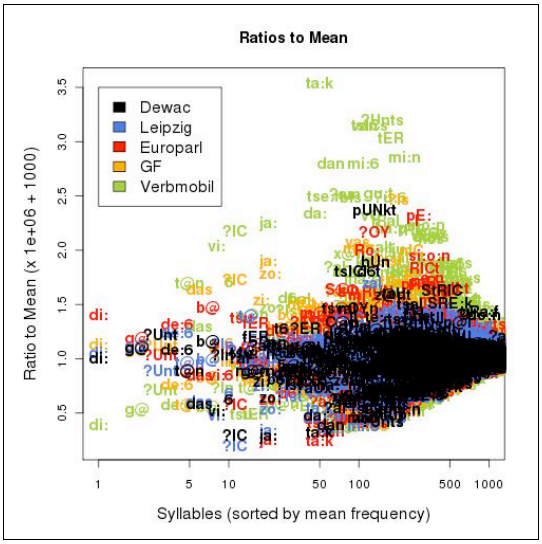

Figure 3: Ratios of syllable frequencies per million in five corpora to their means (adding 1000 first) [25]

This method was adapted here to discover characteristic differences in syllable frequencies among the five corpora. Individual syllable frequencies per million syllables were compared to their mean values, adding 1000 as a constant in order to focus on the more frequent types. Figure 3 shows the resulting values together with the syllable transcriptions for the first 1000 types when sorted by mean. Once again, it becomes clear that most of the strongest deviations belong to the Verbmobil corpus, whereas the Leipzig corpus contains the least amount of highly characteristic syllables. The five most strongly overrepresented syllable types from each corpus are listed in Table 1.

Many of the syllables particularly overrepresented in one corpus can be attributed to words that are highly characteristic of their respective corpus domain. The four top syllables from Europarl, for instance, form most of the word "europäische" (European), and the syllables [? U n t s], [t s I C s], and [v a n] as well as the not-so-characteristic, but high-frequent syllable [t @ n] from the Verbmobil database all correspond to the phrase "x-und-zwanzigsten" (twenty-xth), demonstrating its focus on conveying information about dates.

Other syllables are less intuitively understandable, but can still be interpreted in the context of the corpus in which they are overrepresented. The syllable [d @ s] from the Leipzig corpus appears most often in compound words containing the term "Bundes" (federal), and as such shows the prominent role that governmental agencies play in news reports. The syllables [v a s] and [m a n], which are characteristic for the GF database and form part of the word "(irgend)etwas" (something) and the impersonal pronoun "man" (you/one) respectively, can be seen as indicators of a tendency towards vague, abstract statements in conversational speech.

Corpus	Top Five Characteristic Syllables				
Dewac	p U N k t [pʊŋkt]	h U n [hʊn]	t s I C [tsɪç]	d 6 t [dɐt]	S t R I C [ʃtʁɪç]
Leipzig	z aI [zaɪ]	h a [ha]	z a: k [za:k]	d @ s [dəs]	t aU [taʊ]
Europarl	p E: [pɛ:]	I [ɪ]	? OY [ʔɔʏ]	R o: [ʁo:]	s i: o: n [si:o:n]
GF	? I s [ʔɪs]	v a s [vas]	m a n [man]	n I C [nɪç]	j a: [ja:]
Verbmobil	t a: k [ta:k]	? U n t s [ʔʊnts]	t s I C s [tsɪçs]	v a n [van]	t E R [tɛʁ]

Table 1: Five most overrepresented syllables in each corpus, transcribed in SAMPA and IPA (see Figure 3)

Occasionally, there is no obvious reason for the comparatively high frequency of the syllables in question. Although the syllables [h U n], [d 6 t], and [t s I C] in the Dewac corpus have a clear connection with numbers – [h U n] and [d 6 t] make up the German word for "hundred" and [t s I C] is a morpheme marking the "tens" in numbers, similar to the "-ty" in "twenty" or "thirty" – their role in a corpus of internet texts remains uncertain.

Only a few cases of apparent differences between written-language corpora on the one hand and spoken-language corpora are visible in Figure 3. Here, the Europarl corpus tends to resemble the written-language corpora rather than the corpora GF and Verbmobil. Syllables which are untypical of written language include the affirmative [j a:] ("ja"), which can also be used as a filler and a feedback particle, as well as the first person pronouns [? I C] ("ich") and [v i: 6] ("wir"), the last being (incorrectly) split by the tagger into [v i:] and [6]. Two further cases of syllables underrepresented in written language which are recognizable in this depiction are [d a s], an article or demonstrative pronoun, and [z o:], which appears in words such as "so" (as, so, that way), "also" (so), and "sogar" (even). These results match findings from Allwood's study, where the Swedish words for "so", "I", "yes", "one", and "it" proved to have a higher frequency in the spoken-language corpus than in the written-language database [7].

4. Discussion

All three diagrams depict a higher variation for the conversational corpora GF and Verbmobil. Several possible explanations exist for this result. On the one hand, some of the variation can be traced back to systematic vocabulary differences between spontaneous spoken language and written texts. On the other hand, the differences could also be a result of the comparatively small size of the spoken language corpora, which might make them more contingent on the individual topics of the corpus texts used, hence leading to less robust estimates. Furthermore, the corpus with the highest frequency deviation (Verbmobil) is not only the smallest of the five databases analyzed, but also consists of appointment-making dialogues, a restricted domain with its own characteristic vocabulary.

A comparison of Figures 1 and 2 shows that for frequent syllables, rank numbers tend to be more robust than relative

frequency counts. As syllables become rarer, however, rank differences between different corpora increase while differences in actual relative frequency decrease. Among the very low-frequency syllables, many share the same frequency but are assigned different ranks according to alphabetical order or by chance, making ranking information completely arbitrary in terms of frequency differences [26]. In this way, both representations are highly dependent on the frequency of the compared syllables, one potentially overemphasizing and the other underplaying differences across corpora.

The method of computing frequency ratios described in [25] allows the focus to be placed on different frequency ranges by choosing the added constant accordingly. An examination of frequent syllables using this technique showed only a few instances where differences could clearly be attributed to either spoken or written language. There are several reasons why differences between individual corpora might be more conspicuous than differences between spoken and written language. Because the transcriptions are created automatically from orthographic texts, they are very broad and follow the canonical pronunciation. Therefore, potential differences due to coarticulation effects or hesitation noises go unnoticed. Also, the corpora of spoken and written language do not necessarily have the same degree of formality: Europarl features a highly planned, formal speaking style, while Dewac includes elements of informal, conversational language from forum or chat discussions. And finally, the overrepresented syllables often form part of different inflections and derivations of the same key word, maximizing its effect.

5. Conclusion

By means of automatic phonetic transcription and syllable tagging it is possible to analyze syllable frequencies from written corpora as well as from orthographic transcriptions of spoken-language corpora. A direct comparison of relative syllable frequencies in various databases shows considerable differences among common syllables, most noticeable in corpora of conversational speech. In spite of these differences, syllable ranking remains rather robust. Although there are a few instances where differences in frequency can clearly be attributed to the language modality, the majority of strongly overrepresented syllables in one corpus compared with the others is due to key words characteristically used in the particular corpus domain. The results of the present study indicate that, barring a few exceptions, syllable frequencies from written corpora can be taken as a rough estimate of their frequency in spoken language.

6. Acknowledgements

This study was conducted as part of the project "The syllable as a processing unit in speech production: Evidence from frequency effects on coarticulation", funded by the German Research Foundation DFG (Priority Program 1234).

7. References

[1] W. J. M. Levelt and L. Wheeldon, "Do speakers have access to a mental syllabary?," *Cognition,* vol. 50, pp. 239-269, 1994.

[2] J. Cholin, W. J. M. Levelt, and N. O. Schiller, "Effects of syllable frequency in speech production," *Cognition,* vol. 99, pp. 205-235, 2006.

[3] U. Benner, I. Flechsig, G. Dogil, and B. Möbius, "Coarticulatory resistance in a mental syllabary," in *Proceedings of the International Congress of Phonetic Sciences*, Saarbrücken, 2007, pp. 485-588.

[4] B. Möbius, "Rare events and closed domains: Two delicate concepts in speech synthesis," *International Journal of Speech Technology,* vol. 6, pp. 57-71, 2001.

[5] A. Schweitzer and B. Möbius, "Exemplar-based production of prosody: Evidence from segment and syllable durations," in *Proceedings of the Speech Prosody 2004*, Nara, 2004, pp. 459-462.

[6] M. A. K. Halliday, "The spoken language corpus: A foundation for grammatical theory," in *Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23) Göteborg 22-26 May 2002*, K. Ajmer and B. Altenberg, Eds.: Rodopi, 2002, pp. 11-38.

[7] J. Allwood, "Some frequency based differences between spoken and written Swedish," in *Proceedings of the 16th Scandinavian Conference of Linguistics*, Turku, 1998.

[8] W. Chafe and D. Tannen, "The relation between written and spoken language," *Annual Review of Anthropology,* vol. 16, pp. 383-407, 1987.

[9] F. N. Akinnaso, "On the similarities between spoken and written language," *Language and Speech,* vol. 28, p. 323, 1985.

[10] M. Baroni and A. Kilgarriff, "Large linguistically-processed web corpora for multiple languages," in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, 2006, pp. 87-90.

[11] U. Quasthoff, M. Richter, and C. Biemann, "Corpus portal for search in monolingual corpora," in *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genua, 2006, pp. 1799-1802.

[12] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of the 10th Machine Translation Summit*, Phuket, 2005, pp. 79-86.

[13] S. Dickgießer. (2004). Korpus GF, Gespräche Im Fernsehen: Talkshows, Diskussionen, Interviews. [Online]. Available: http://agd.ids-mannheim.de/html/korpora/pdf/gfdok.pdf.

[14] W. Wahlster, "Mobile speech-to-speech translation of spontaneous dialogs: An overview of the final Verbmobil system," in *Verbmobil: Foundations of speech-to-speech translation*, W. Wahlster, Ed. Berlin: Springer, 2000, pp. 3-21.

[15] P. A. Taylor, *Text-to-speech synthesis*. Cambridge; New York: Cambridge University Press, 2009.

[16] IMS Uni Stuttgart. Speech Synthesis at the IMS - Download IMS German Festival. [Online]. Available: http://www.ims.uni-stuttgart.de/phonetik/synthesis/festival_opensource.html.

[17] T. Portele, J. Krämer, and D. Stock, "Symbolverarbeitung im Sprachsynthesesystem Hadifix," in *Proceedings of the 6. Konferenz Elektronische Sprachsignalverarbeitung*, Wolfenbüttel, 1995, pp. 97-104.

[18] J. C. Wells, "SAMPA - computer readable phonetic alphabet," in *Handbook of standards and resources for spoken language systems*. Berlin; New York: Mouton de Gruiter, 1997, pp. 684-732.

[19] H. Schmid, B. Möbius, and J. Weidenkaff, "Tagging syllable boundaries with joint n-gram models," in *Proceedings of the Interspeech 2007*, Antwerpen, 2007.

[20] SK Uni Bonn. Bomp - ein maschinenlesbares deutsches Aussprachewörterbuch. [Online]. Available: http://www.sk.uni-bonn.de/forschung/phonetik/sprachsynthese/bomp.

[21] R. H. Baayen, *Word frequency distributions*. Dordrecht; Boston: Kluwer Academic, 2001.

[22] M. Baroni, "Distributions in text," in *Corpus linguistics: An international handbook,* vol. 2, A. Lüdeling and M. Kytö, Eds. Berlin: Mouton de Gruyter, 2009, pp. 803-821.

[23] R Development Core Team. (2010). R: A language and environment for statistical computing. [Online]. Available: http://www.r-project.org.

[24] S. Evert and M. Baroni, "zipfR: Word frequency distributions in R," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, 2007, pp. 29-32.

[25] A. Kilgarriff, "Simple maths for keywords," in *Corpus Linguistics Conference*, Liverpool, 2009.

[26] A. Kilgarriff, "Comparing corpora," *International Journal of Corpus Linguistics,* vol. 6, p. 97, 2001.