A PHONOLOGICAL DATABASE BASED ON CELEX AND UNIGRAM FREQUENCIES

Cyrus Shaoul & Fabian Tomaschek

Department of General Linguistics, University of Tübingen, Germany

A PHONOLOGICAL DATABASE BASED ON CELEX AND UNIGRAM FREQUENCIES

The phonological database provides a large number of word and phoneme counts. It merges two currently well established databases: CELEX and SDEWAC into a user-friendly R datafile.

Corresponding author:

Fabian Tomaschek

Seminar für Sprachwissenschaft

Eberhard Karls University Tübingen

Wilhelmstrasse 19

Tübingen

e-mail: fabian.tomaschek@uni-tuebingen.de

Version: October 9, 2015

0pt

Introduction

In recent years phonetic studies have concentrated on measuring linguistic experience on the phonetic signal. In order to perform such investigations large word and phonemic corpora are needed. For such a purpose the present database has been created. The phonological database is a merge between CELEX and unigram frequencies from the SDEWAC corpus (Faaß & Eckart, 2013). It is still work in progress but nevertheless enables the user to find a large number of measurements for a variaty of environments.

Usage

In order to use the database, you need to have installed "R" on your system.

In order to check the structure of the variables, type `str(VarName)`.

Files

Unigram Frequencies

You can load the database by typing into R:

```
load('YOURDIRECTORY/Unigram_Frequencies.rda')
```

Items in the variable:

- **Word:** The word in question.
- **RawFreq:** The count in den SDEWAC corpus.
- **FreqPerMillion:** The number of occurences per million in the SDEWAC corpus.
- **NSyl:** Number of syllables in the word, counted by the number of vowels.
- **VokPos:** Location of the vowel in the word.

- **Upper:** Indication whether the word beginns with an upper-case letter. 0 = no, 1 = yes. Concretely this can be used in order to find German nouns. However, it will also show all words wich are located at the beginning of a sentence (e.g. 'Aber' and 'aber').
- **logFreq:** The natural logarithm of FreqPerMillion. Usually used for analyses because counts are not normally distributed.
- **Syl.X:** Indicates, what vowel is located in the first, second, etc. syllable. However, this is based on the orthographical representation, i.e. <a,e,i,o,u,ü,ä>. I.e., No schwa is coded. Diphthongs are coded as numbers: <au> = 1, <eu> = 2, <ei> = 3.
- **Str:** Vowel-consonant structure of the word.

Bigram Frequencies

You can load the database by typing into R:

```
load('YOURDIRECTORY/Bigram_Frequencies.rda')
```

Celex Phonology Lemma

You can load the database by typing into R:

```
load('YOURDIRECTORY/Celex_Phonology_lemma.rda')
```

Unfortunately, I dont have any recent lemma frequencies, as these are hard to compute. If you know a good lemma frequency corpus, just let me know and I will include it.

Celex Phonology Wordforms

You can load the database by typing into R:

```
load('YOURDIRECTORY/Celex_Phonology_wordforms.rda')
```

The frequencies in this database are taken from the wordform frequencies in "Unigram_Frequencies"

Phonemic distributions and probabilities

You can load the database by typing into R:

```
load('YOURDIRECTORY/Corp_01_phonemic_distributions_probabilities_activations.rda')
```

TG.abc TG.ab. These two variables contain NDL measurements. NDL calculates the activation weight between an outcome and its cues. Read more about NDL activations in Baayen, Milin, Đurđević, Hendrix, and Marelli (2011).

Activations: Activations indicate how strong an outcome is activated by a given cue combination.

MAD: MADs specify how big is the activation deviation from the median for a certain outcome given the activations of all cue combinations for that outcome.

segment_monograms/bigrams/trigrams. These three variables contain segment and segment sequences counts. Counts were performed by looking up the all words in which a specific segment/sequence occurs and summing up the word's frequencies in Unigram_Frequencies. The columns specify:

- **A,B,C:** A, B, C, seq: part of the segment/sequence
- **N:** Raw count
- **P:** Probability of the segment/sequences (P(x) = N(x)/sum(x1 ... xn))
- **cndP.ab_a:** Conditional probability of B given A (calculated by: P(b|a) = P(b)/P(ab), thus the name)
- **cndP.ab_b:** Inverse conditional probability of A given B (calculated by: P(a|b) = P(a)/P(ab), thus the name)
- **Hfam.a_b:** Entropy in the B position given A. (calculated by: -1* sum(D*log2(D)), where D is a vector P(AB1 ... ABn). Entropy tells you how easy (=low values) or hard (=high values) it is to find B given A.

The same measures were used in the segment_trigrams variable.

Download

https://www.dropbox.com/sh/j8td1ro6ptjpvr1/LkomVE47Gr

Citation

If you use the database, please cite it as: Shaoul, C., Tomaschek, F. (2013): A phonological database based on CELEX and N-gram frequencies from the SDEWAC corpus. Personal communication.

Bibtex entry:

```
 @ARTICLE{PDB:ST,
Author = "Cyrus Shaoul and Fabian Tomaschek",
Title = "A phonological database based on CELEX and N-gram
frequencies from the SDEWAC corpus",
Year = "2013",
Journal = "Personal communication"
}
```

Disclaimer and Contact

As always, errors and typos can occur when you work with databases as large as this. Please let me know if you find anything. Furthermore, the current database is work in progress and will be changed constantly. Also, do not hesitate to contact me, if you have specific questions. Please do not pass this database on to third parties without our approval.

fabian.tomaschek@uni-tuebingen.de

Dr. Fabian Tomaschek

Quantitative Linguistics

University of Tübingen

Wilhelmstraße 17

72074 Tübingen

fabian.tomaschek@uni-tuebingen.de

References

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review*, *118*(3), 438–481.

Faaß, G. & Eckart, K. (2013). Sdewac – a corpus of parsable sentences from the web. In I. Gurevych, C. Biemann, & T. Zesch (Eds.), *Language processing and knowledge in the web* (Vol. 8105, pp. 61–68). Lecture Notes in Computer Science. Springer Berlin Heidelberg. doi:10.1007/978-3-642-40722-2_6