# Swedish Word Metrics: A Swe-Clarin resource for psycholinguistic research in the Swedish language

**Erik Witte**
Swedish Institute for Disability Research
Linköping University
Linköping, Sweden
erik.witte@liu.se

**Jens Edlund**
Speech, Music and Hearing
Royal Institute of Technology
Stockholm, Sweden
edlund@speech.kth.se

**Arne Jönsson**
Computer and Information Science
Linköping University
Linköping, Sweden
arne.jonsson@liu.se

**Henrik Danielsson**
Swedish Institute for Disability Research
Linköping University
Linköping, Sweden
henrik.danielsson@liu.se

**Abstract**

We present *Swedish Word Metrics* (SWM), a new CLARIN resource for calculations of lexical and sub-lexical metrics of Swedish words. The calculations at SWM are based on the AFC-list, which is a freely available lexical database with 816404 entries containing spellings, phonetic transcriptions, word-class assignments, and word frequency data. Besides allowing for easy access to the AFC-list data, the SWM site calculates metrics of orthographic and phonological neighbourhood density, phonotactic probability, orthographic transparency, as well as phonetic and orthographic isolation points. The source code for all calculations has been made publicly available and can be extended with more types of word metrics, whereby it forms a framework for continued word-metric developments in the Swedish language.

## 1 Introduction

Over the years, researchers within the field of psycholinguistics have noted that certain lexical and sublexical properties systematically impact the process of human word perception. Most prominently, the *word-frequency effect* cause words with high word frequency (WF), i.e. that often occur in spoken or written language, to be more quickly and accurately perceived than less frequently occurring words (Brysbaert et al., 2018).

Alongside WF, also other metrics are important for word recognition. Of these, the effect of neighbourhood density (ND) (Luce and Pisoni, 1998) have been extensively studied. The neighbourhood of a word is typically considered to consist of other similar words and described on a scale ranging from sparse neighbourhoods, with only a few neighbours, to dense neighbourhoods containing many similar words. In the auditory domain, phonologically similar words appear to compete with each other, making the process of word recognition more difficult in dense neighbourhoods compared to sparse. However, when orthographic neighbourhoods are considered, studies have indicated a reversed effect, making spoken high-density words easier to perceive than low-density words. Ziegler et al. (2003) speculated that this could be related to irregularities in how words are represented orthographically. Several studies have since shown that the level of orthographic transparency (OT) influences, not only the reading process, but also word recognition in the auditory domain (Dich, 2014). Also, the likelihood of encountering specific phonemes in different parts of words, a property often referred to as phonotactic probability (PP), seems to have a facilitatory effect upon word recognition (Vitevitch and Luce, 1999).

## 2 The AFC-list

Forming a base resource for calculations of psycholinguistic metrics in the Swedish language, the lexical database referred to as the AFC-list (Witte and Köbler, 2019) contains spellings, phonetic transcriptions, word-class assignments and WF data drawn from several freely available sources. The AFC-list contains 816404 entries, each constituting a unique combination of phonetic transcription and spelling. The distribution of the number of syllables per word in the AFC-list is presented in the leftmost pane of Figure 1.

Based on the AFC-list, Witte and Köbler (2019) defined Swedish versions of word metrics previously published for other languages — including neighbourhood density, orthographic transparency and phonotactic probability — and also developed several new word-metric algorithms specifically targeted towards Swedish phonology. Finally, Witte and Köbler (2019) calculated their metrics for all entries in the AFC-list and made the results publicly available in supplementary data files.

## 3 The Swedish Word Metrics website

The purpose of the current study was to make calculations of Swedish word metrics, as well as the content of the AFC-list, directly available via an internet website.

Towards this aim, we have created the Swedish Word Metrics[1] (SWM) website, which implements all word-metric calculations described by Witte and Köbler (2019) as well as functionality for searching directly in the AFC-list. In addition to the word metrics described in Witte and Köbler (2019), we have also implemented a few additional word metrics such as phonetic and orthographic *isolation points* (Marslen-Wilson and Welsh, 1978), as well as the *Orthographic Levenshtein Distance 20* (OLD20) developed by Yarkoni et al. (2008) into the SWM website. The SWM website is hosted by the Swedish national research infrastructure *Språkbanken Tal*, which is a member of Swe-Clarin, the Swedish node in the European CLARIN research infrastructure for language resources and technology.

The word-metric calculation functionality is found on the SWM website's *Calculator* page. Words to input into the calculator should be specified by their spellings, their phonetic transcriptions, and an optional word frequency value. All phonetic transcriptions need to adhere to the transcription convention described in Witte and Köbler (2019). To this aim, the calculator contains an optional transcription checking facility that performs phonological evaluations of the entered phonetic transcriptions. For words already present in the AFC-list, it is sufficient only to supply the spelling; all other data will be drawn automatically from the AFC-list. The calculator interface also allows for customisation of which word-metric calculations to run. This feature is useful since some metrics — for instance ND — can be quite time consuming to calculate.

In the SWM website's *AFC-list* page, the AFC-list can be searched either by specifying any combination of spelling, phonetic transcription and word frequency, or by using a SQL query, enabling full search term flexibility.

Both the word-metric calculation results and the AFC-list search results are returned in tables in which rows represent entered words or matching AFC-list entries and columns present the different word metrics and other descriptive properties of each entry. The tables contain up to a total of 70 columns, all described on the SWM website's *Info* page.

Below we will briefly describe a selection of the lexical and sub-lexical word metrics that can be calculated at the SWM website.

### 3.1 Frequency-weighted phonological neighbourhood density

Besides raw WF data, the AFC-list also contains the *Zipf-scale* value of each included word. The Zipf-scale is a WF metric developed for the specific purpose of capturing the WF effect upon word recognition (van Heuven et al., 2014). The Zipf-scale value takes into account both the

[1] https://www.sprakbanken.speech.kth.se/data/swm/

total size of the corpus from which the frequency data was derived as well as the number of word types in that corpus. In addition, the Zipf-scale value is constructed to assume that there exists a number of words in the language which are not included in the corpus used. The Zipf-scale values were used by Witte and Köbler (2019) to create a neighbourhood metric that takes both WF and ND into account. The metric was referred to as the Zipf-scale weighted phonetic neighbourhood density probability (PNDP) and expresses the Zipf-scale weighted probability of encountering a specific word given the Zipf-scale values of its phonological neighbours (Witte and Köbler, 2019). The SWM calculator identifies phonological neighbours as other existing words which differ from the target word by an edit distance (i.e., one insertion, deletion or substitution) of one, and share the same number of syllables. The comparison words are primarily taken from the AFC-list, but an option in the SWM calculator also allows manually entered words not present in the AFC-list to be included in the comparisons.

To illustrate, the words *bladet* [blɑːdẹt] (the sheet) and *floder* [flʉːdẹr] (rivers) are very similar in terms of syllabic structure, WF, OT and PP, but differ in their PNDP values (0.16 and 0.78, respectively) since *bladet* has several, more common, phonological neighbours (e.g. *badet* [bɑːdẹt] (the bath), *blodet* [blʉːdẹt] (the blood), *bladen* [blɑːdẹn] (the leaves)) while *floder* has only one, less common, neighbour (*floders* [flʉːdẹʂ] (rivers')).

### 3.2 Orthographic transparency

In order to calculate metrics of OT, specific segments of the phonetic transcriptions, here referred to as pronunciations, need to be matched to their corresponding segments in the spelling, i.e. graphemes. In the AFC-list, such grapheme/pronunciation correspondences are called *sonographs*. The SWM calculator determines the sonographs of each entered word using a rule-based parsing algorithm implementing a custom-made finite-state transducer of the same type as described in Witte and Köbler (2019). To exemplify, it parses the sonographs in the word *sakens* [sɑːkẹns] (the thing's) and the word *djungeln* [j̧øŋːẹln] (the jungle) into (s-s|a-ɑː|k-k|e-ẹ|n-n|s-s) and (dj-j̧|u-ø|ng-ŋː|e-ẹ|l-l|n-n), respectively.

Based on the AFC-list sonographs, Witte and Köbler (2019) determined word-specific probabilities for each grapheme to correspond to different pronunciations, here called *grapheme-to-pronunciation* (G2P)-OT, according to the method developed by Berndt et al. (1987). Since basing OT calculations directly upon grapheme-pronunciation correspondences makes the unmerited assumption that readers know the length of each encountered grapheme before parsing it. Therefore, Witte and Köbler (2019) also developed a modified OT metric referred to as the *grapheme-initial letter-to-pronunciation* OT (GIL2P-OT) in which OT accounts for both the process of identifying a grapheme given its first letter and the process of finding the appropriate pronunciation for that grapheme. The SWM calculator uses the probability data calculated in Witte and Köbler (2019) to calculate all three types of OT metrics. While for the example words *sakens* and *djungeln* given above, the WF, ND and PP values returned from the SWM calculator are very similar, the word-average GIL2P-OT values are relatively different (0.99 and 0.92, respectively). This difference is primarily related to the unusual situation in which a grapheme-initial letter *d* initiates the pronunciation [j̧].

### 3.3 Phonotactic probability

Besides the commonly used, and linguistically neutral, phonotactic metrics by Vitevitch and Luce (2004), Witte and Köbler (2019) introduced a new metric which determines PP separately for different types of syllables as well as different intra-syllabic positions. This metric is referred to as the *normalised stress and syllable structure-based* PP (SSPP). To illustrate, the SWM calculator output for the words *skrattet* [skratːẹt] (the laughter) and *sniglar* [sniːglar] (snails) are very similar in WF, ND and OT, but differ largely in their SSPP values (0.97 and 0.89, respectively). The reason for this difference is the relatively rare occurrences of the bi-phones [sn] in syllable onsets, [iːg] in the rhyme, and [gl] across a syllable boundary.

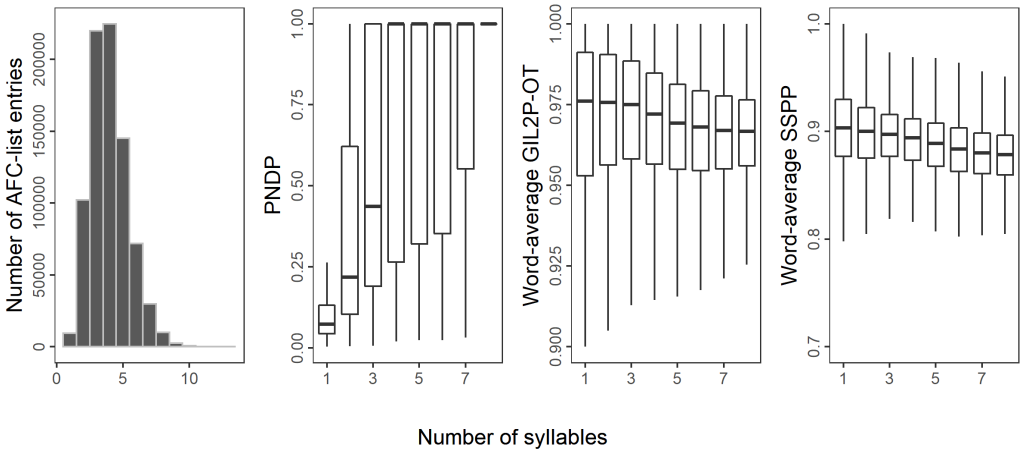


Figure 1. From left to right, a histogram presenting the word lengths (in number of syllables) of all entries in the AFC-list, and boxplots (outliers removed) presenting the distributions of Zipf-scale weighted phonetic neighbourhood density probability (PNDP) values, grapheme-initial letter-to-pronunciation orthographic transparency (GIL2P-OT) values, and word-average normalised stress and syllable structure-based phonotactic probability (SSPP) values for different word lengths in the AFC-list, respectively.

Figure 1 presents the distribution of PNDP, word-average GIL2P-OT and SSPP for different word lengths in the AFC-list. As is clearly seen, PNDP is lowest for short words and increases rapidly with increasing word lengths. PIP2G-OT and SSPP values are relatively stable across word lengths, both showing a slight decrease towards longer words.

## 4 Potential applications

Since most metrics in the SWM resource were developed with the aim to reflect and quantify different psycholinguistic phenomena, they can potentially be used to model different aspects of human word perception computationally. The large size of the AFC-list makes it an appropriate resource in the creation of experiments by which hypotheses stemming from such models can be tested, thus advancing current theories about human speech perception. Potentially, word metrics such as ND, OT and PP can also be used to improve the quality of synthetic speech. For example, the metrics could be utilised to adjust the local playback speed of speech-synthesis algorithms to assimilate natural variations in reading speed caused by the level of orthographic and phonotactic complexity of specific words, or by the presence of competing phonological neighbours. Furthermore, as OT has bearings upon the ease of phonological decoding of written words, it could likely be used to improve the predictive accuracy of readability metrics such as the commonly used Swedish LIX or other similar metrics (see Heimann Mühlenbock, 2013).

## 5 Source-code availability

The word-metric calculations used by the SWM website were written in an independent and cross-platform software library using .NET. This backend library is called *Swedish Word Metrics Calculations* (SWMC) and its source code has been made publicly available under an Apache-2.0 license via a link at the SWM website.

## References

Berndt, R. S., Reggia, J. A., and Mitchum, C. C. 1987. Empirically Derived Probabilities for Grapheme-to-Phoneme Correspondences in English. *Behavior Research Methods, Instruments, & Computers*, 19(1):1–9.

Brysbaert, M., Mandera, P., and Keuleers, E. 2018. The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*, 27(1):45–50.

Dich, N. 2014. Orthographic Consistency Affects Spoken Word Recognition at Different Grain-Sizes. *Journal of Psycholinguistic Research*, 43(2):141–148.

Heimann Mühlenbock, K. 2013. *I See What You Mean. Assessing Readability for Specific Target Groups.* Doctoral dissertation, University of Gothenburg, Gothenburg, Sweden.

Luce, P. A. and Pisoni, D. B. 1998. Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, 19(1):1–36.

Marslen-Wilson, W. D. and Welsh, A. 1978. Processing Interactions and Lexical Access During Word Recognition in Continuous Speech. *Cognitive Psychology*, 10(1):29–63.

van Heuven, W. J. B., Mandera, P., Keuleers, E., and Brysbaert, M. 2014. SUBTLEX-UK: a New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*, 67(6):1176–1190.

Vitevitch, M. S. and Luce, P. A. 1999. Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition. *Journal of Memory and Language*, 40(3):374–408.

Vitevitch, M. S. and Luce, P. A. 2004. A Web-Based Interface to Calculate Phonotactic Probability for Words and Nonwords in English. *Behavior Research Methods, Instruments, & Computers*, 36(3):481–487.

Witte, E. and Köbler, S. 2019. Linguistic Materials and Metrics for the Creation of Well-Controlled Swedish Speech Perception Tests. *Journal of Speech, Language, and Hearing Research*, 62(7):2280–2294.

Yarkoni, T., Balota, D., and Yap, M. 2008. Moving Beyond Coltheart's N: a New Measure of Orthographic Similarity. *Psychonomic Bulletin Review*, 15(5):971–979.

Ziegler, J. C., Muneaux, M., and Grainger, J. 2003. Neighborhood Effects in Auditory Word Recognition: Phonological Competition and Orthographic Facilitation. *Journal of Memory and Language*, 48(4):779–793.