# Using lexical and corpus resources for augmenting the AAC lexicon

**Katarina Heimann Mühlenbock**
Dept of Swedish Language
University of Gothenburg
Gothenburg
katarina.heimann.muhlenbock@gu.se

**Mats Lundälv**
Dart
Queen Silvia Children's Hospital
Gothenburg
mats.lundalv@vgregion.se

## Abstract

A corpus of easy-to-read texts in combination with a base vocabulary pool for Swedish was used in order to build a basic vocabulary. The coverage of these entries by symbols in an existing AAC database was then assessed. We finally suggest a method for enriching the expressive power of the AAC language by combining existing symbols and in this way illustrate additional concepts.

## 1 Introduction

A considerable proportion of the population, among 1.3 % of all individuals (Beukelman and Mirenda, 2005) are affected by severe communication disorders, making them more or less unable to use written and/or spoken language. Different language supportive aids for these persons have evolved over the years, mainly as graphical systems containing symbols and pictures, simplified supportive signing (derived from sign language vocabulary), or a combination of these, possibly comprising speech synthesis and speech recognition. All these supportive measures and methods are referred to as Augmentative and Alternative Communication (AAC).

A vocabulary comprising 20,878 lemma or base forms from different sources was analysed in terms of frequency and dispersion. The primary issue in this study was to analyse to what extent the concepts in the main AAC symbol databases mirror the vocabulary needed to produce and understand everyday Swedish language. Another goal was to investigate the possibility of extending the AAC symbol databases by combining separate basic words from the vocabulary into compounds.

## 2 Background

A fundamental aspect for participation in the society is the possibility to acquire information and to communicate. For the majority of citizens, getting information on every-day issues is hardly a task entailing any specific problems. There is, however, a substantial number of persons who have substantial difficulties to benefit from ordinary written and spoken sources, being dependent upon other modalities, either to express themselves, or as a support for interpretation, or both. For this purpose, AAC resources in the shape of pictures and symbols have been designed for use in low-tech solutions such as communication books and boards, and high-level technologies such as computer programs and eye-tracking devices. AAC resources at hand are, however, burdened by two major problems. First, manual production requires much effort and time. New concepts have to be identified and a uniform agreement has to be reached among different parties on how to visually present the concept. Second, accessibility in the sense of a consumer's or developer's possibility and freedom to use available resources, is strongly restricted by distribution, copyright and licensing issues. Different projects have been carried out with the goal to develop and implement some suggested open standards for syntactic and semantic encoding of AAC material. The European IST project WWAAC (World Wide Augmentative & Alternative Communication, 2004) was



*Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pages 120–127,
Edinburgh, Scotland, UK, July 30, 2011. 

a pan-european initiative to make the web more accessible for a wide range of persons with language and/or cognitive impairments.

An essential part of language lies within its ambiguity on the lexical as well as structural level. When it comes to computerized processing, semantic variation between word forms, morphological relationships within different word forms, and multiword items claim specific handling, especially when enriching an existing vocabulary with new entries. In fact, comparing wordlists and frequencies from different sources is a task affected by a couple of complications. One problem encountered in a comparative study of word frequencies is how a *word* is defined, which in fact has been put under debate by for instance Gardner (2007). In the present study, we consider the *lemma*, i.e. the look up form of a word, to be the key unit. The idea behind the use of lemma frequencies as units of study is that the human mental or computational processing of lemmas and inflected forms profit from each other, which is in favour of a theory implying that a morphological decomposition takes place for the recognition of inflected forms.

Knowledge of the vocabulary is an essential part of both conveying and understanding messages, verbally as well as non-verbally. Together with the system of rules generating grammatical combinations, the words in the vocabulary contribute to the infinite expressive power of language. With a narrow vocabulary, the possible means to produce and achieve adequate information decreases. Researchers have attempted to identify lists of words that could be included in a core vocabulary (Thorndike and Lorge, 1944), (Spache, March 1953) and more specifically for people who use AAC (Balandin and Iacono, 1998), (Banajee et al., 2003). There have also been efforts to investigate how much of the vocabulary a person needs to know in order to grasp the content of written texts without having to struggle with isolated, unknown words (Hirsch and Nation, 1992). In the latter study, a list of around 2,000 high frequency words of English, compiled by West (1953), was used in order to investigate if knowledge of these words was actually sufficient for reading unsimplified short novels. It was found that a person with this restricted vocabulary was familiar with about 90-92% of the total words. It is worth noting that the word frequency counts here reflect the number of times a word pertaining to a certain *word family* occurs in a text. The idea behind a word family is that inflected and regulary derived forms of a known base word can also be considered as known words if the affixes are known. This implies that nouns, adverbs, verbs and adjectives sharing a common base will be treated as one word, contrary to the lexicographical traditions (for Swedish), where the lemma or base form is the conventional unit of study.

With this in mind, it follows that a database containing a core vocabulary of a language has to contain enough information for identification of different lexemes. For our purposes in this study, it was also necessary to add another source of information in order to retrieve the semantic identifiers for subsequent illustration of different concepts into AAC symbols.

## 3 Experimental design

A corpus of easy-to-read texts and children's fiction was used in order to retrieve candidates for inclusion into a database of basic Swedish. The hypothesis is that by using a corpus of plain texts produced with the aim of being easy to understand, we can derive appropriate data for further statistical analysis of which words or word forms are to be considered as pertaining to a basic vocabulary. The candidates retrieved by excerption of high-frequency lemmas from the corpus were subsequently compared to the base-form words in a Swedish base vocabulary, where the lemmas obtaining the highest rank in both sets were integrated into a database of core vocabulary. The AAC symbol coverage of these database entries was then assessed by addressing an existing AAC symbol database. Finally, attempts were made to expand the existing AAC vocabulary through a semantic analysis of new words, simple as well as compounds, and in that way make it possible to illustrate new concepts.

## 4 Material

The material used comprise corpora as well as lexica. Some of the resources are freely available from the public domain, while other are used under specific permissions.

### 4.1 AAC material

Pictures and symbols aiding language and communication have been developed over decades. Some of the symbol systems have a visual structure that supports different parts of speech. For this study, the Widgit symbols library (Widgit Software, 2011) and vocabulary in Swedish (preliminary version 11) was used, covering around 11,000 symbols and 64,000 words (including synonyms and inflected forms). Some of the symbols are produced in order to illustrate different concepts rather than isolated words, which to some extent had a negative impact on the comparison of different wordlists. The focus of interest has been on content words, i.e. nouns, verbs, adjectives and adverbs, since the functional words normally don't appear as independent items. In total, a wordlist of 20,907 entries was extracted, normally the lemma form. Proper nouns and numbers were excluded in the study.

### 4.2 Corpora

#### 4.2.1 LäSBarT

The primary corpus material for this study is LäSBarT, an acronym for *Lättläst Svenska och Barnbokstext* 'Easy-to-read Swedish and Children's fiction Texts' (Mühlenbock, 2008). It is a specialized corpus of 1.3 million tokens, compiled with the objective to mirror simple vocabulary and syntax. The main text types include works from different domains and genres, such as fiction, official documents from the government, parliament, county council, municipality and daily news. The common denominator for all the texts is that they are all intended to be read by persons that do not fully master everyday Swedish language.

The size of the corpus, 1.3 million tokens, was compensated for by making text representativeness be decisive during compilation. The supply of easy-to-read material is limited and subsequently, the variation range is quite narrow. Contrary to many other writing tasks, the production of easy-to-read text is elicited by a specific need from the society and we cannot expect a large variety of genres. Three genres of easy-to-read texts were identified for obtaining a representative sample, namely fiction, news and community information, which for the target group of readers can be regarded as being a balanced corpus.

#### 4.2.2 SUC 2.0

SUC 2.0 is a balanced corpus of 1 million words in written Swedish, originating from the 1990's. It is designed according to the Brown corpus (Francis and Kucera, 1979) and LOB corpus (Johansson et al., 1978) principles, which means that it consists of 500 samples of text with a length of about 2,000 words each. The state-of-the-art markup language at the time of compilation was SGML, and this annotation schema is kept also in the acctual, revised version. All entries are annotated with parts-of-speech, morphological analysis and lemma, or rather base form. The corpus is also provided with a wide range of structural tags and functionally interpreted tags, according to the TEI standards Sperberg, (Consortium, TEI, 2007).

At the lexeme level, about 23% of the SUC corpus is covered by nouns, while verbs amounts to 17%, adjectives to 9%, proper nouns to 4%, adverbs to 9%, prepositions 12%, conjunctions 8%, numbers 2%, pronouns 10% and determiners to 6% of the total words. The total vocabulary has 69,371 base forms.

### 4.3 Lexica

#### 4.3.1 LäSBarT wordlist

The wordlist obtained from the LäSBarT corpus, *LäSBarT-listan* (henceforward referred to as LBL) contains 22,041 lemmas in total, covering 43,364 lexemes, proper nouns excluded. It contains information about lexical frequency, baseform, part-of-speech tag, and lemma/lexeme form. The lemma/lexeme information tells us that a word like *sticka* has three different lemma/lexeme forms, namely **sticka.1** for the noun *sticka* 'splinter; knitting needle', and **sticka.2** or **sticka.3** for the two different verb lexemes with the meanings 'prick, sting; put' and 'knit', respectively. This information is necessary for further semantic disambiguation of polysemous words.

The overall part-of-speech distribution is listed in Table 1. In this study, 2,277 verbs, 14,856 nouns, 2,715 adjectives and 1,030 adverbs were extracted for further analysis.

| Part-of-speech | % lemmas | % lexemes |
|---|---|---|
| Nouns | 67.4 | 20.1 |
| Verbs | 10.1 | 25.3 |
| Adjectives | 12.3 | 6.2 |
| Adverbs | 4.7 | 10.0 |
| Prepositions | 0.4 | 17.4 |
| Conjunctions | 0.1 | 7.1 |
| Pronouns | 1.2 | 12.6 |
| Determiners | 0.1 | 4.0 |

Table 1: POS-distribution in LBL

It is interesting to note a large discrepancy in verbal representation between SUC (17 %) and LäSBarT (25 %). The most probable explanation to this is the tendency among authors of easy-to-read texts to paraphrase a complicated sentence by two or more simpler ones, each necessitating a new head verb.

#### 4.3.2 The Swedish Base Vocabulary Pool

The Swedish base lemma vocabulary pool (henceforward referred to as SBV) (Forsbom, 2006) is derived from the SUC corpus. The units of the SBV are the base forms from SUC annotation disambiguated for part-of-speech. This means for example that a polysemous and homonymous word pertaining to different parts-of-speech such as a noun and a verb is represented both as its nominal and its verbal form. No information is, however, given at the lexeme or semantic level. The version presently used contains 8,215 entries, where the lemmas are ranked according to relative frequency weighted with dispersion, i.e. how evenly spread-out they are across the subcorpora. Instead of using frequency alone, the formula for adjusted frequency calculation was used (Rosengren, 1972):

$$AF = (\sum_{i=1}^{n} \sqrt{d_i x_i})^2$$

where
$AF$ = adjusted frequency
$d_i$ = relative size of category $i$
$x_i$ = frequency in category $i$
$n$ = number of categories

The SBV was used as reference material for the comparison of dispersion of word base forms to LäSBarT.

#### 4.3.3 SALDO

SALDO (Borin and Forsberg, 2009) is a modern Swedish semantic and morphological lexicon. The organization differs in a fundamental way from the widely used lexical-semantic database Princeton WordNet (Fellbaum, 1998), even though both are based on psycholinguistic principles. While Princeton WordNet and its descendant Swedish WordNet (Viberg et al., 2002), are organized in encoded concepts in terms of sets of synonyms, called synsets, the associative relations between the entries in SALDO are based on metaphorical kinships that are specified as strictly hierarchical structures. Every entry in SALDO must have a mother, which in practice often is either a hyperonym or a synonym. At the top of the hierarchy is an artificial most central entry, the PRIM, which is used as the mother of 50 semantically unrelated entries. In this way, all entries become totally integrated into a single rooted tree without cycles.

## 5 Comparative results

The lemma forms of 2,277 verbs (Fig. 1), 14,856 nouns (Fig. 2), 2,715 adjectives (Fig. 3) and 1,030 adverbs (Fig. 4) in LBL were compared against the SBV in order to obtain lemmas occurring in both lists, i.e. the intersection of two high-frequency and evenly distributed sets of words in the two corpora *LäSBarT* and *SUC*. This yielded a remaining set of 961 verbs, 2,390 nouns, 692 adjectives and 425 adverbs, illustrated as the top two rectangles of each figure. In order to analyse to what extent the AAC symbols really supported this basic vocabulary, an additional comparison was made, focusing on the intersection of words with and without symbol coverage in the two sets. It turned out that as much as 95 % (916 out of 961) of the verbs present in both LBL and SBV also were represented by symbols. For nouns, the corresponding ratio was 76 %, and for adjectives and adverbs 71 % and 60 %, respectively. Figures 1-4 illustrate the overall ratios.

### 5.1 Verbs

Adjusted frequency of the 44 verbs not represented in the symbol database ranged between 14.97 and 1.38, implying a moderate dispersion and frequency. In addition, the majority were compounds with an adverb or preposition as prefix, predominantly composite particle verbs. Authors of easy-to-read texts normally avoid composite particle verbs and prefer to use a paraphrase or synonym, since the former lexical structure can be perceived as old-fashioned and therefore difficult to understand. Furthermore, as many as 29 of the verbs lemmas were hapax words.

Some interesting features must also be mentioned regarding the verbal semantic fields of the words not supported by symbols. Many of the verbs seem to fall into a group of socially motivated actions, such as *bestraffa* 'punish', *fängsla* 'imprison', *beordra* 'command', and *uppfostra* 'educate/rear', all with a rather stern tone.

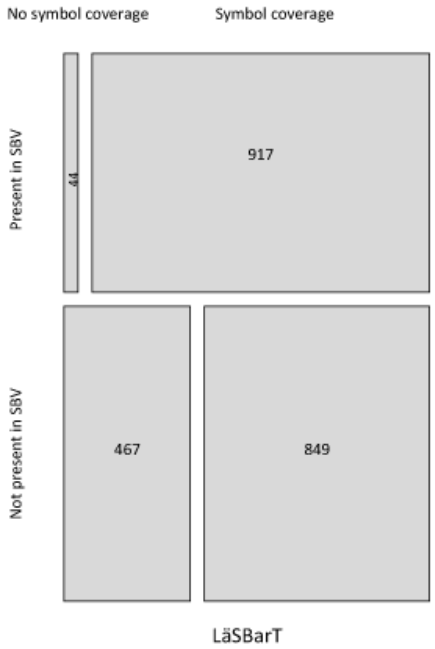


Figure 1: Overall ratio of LäSBarT verbs, presence in SBV and symbol coverage

### 5.2 Nouns

We found that 24 % of the noun lemmas in LBL and SBV lacked symbol coverage, and that there was a wide range in adjusted frequency, varying from 232.84 down to 1.06. Without making any formal categorization, it is clear that the words with highest adjusted frequency are abstract words, such as *samband* 'connection', *brist* 'lack', *sammanhang* 'context', and *allvar* 'seriousness'. Some of the nouns are meaningful only as elements of multiword expressions, such as *skull* 'sake' or *vis* 'manner', while others seem to be ephemeral words from news reports. One third are hapax, and 24 % of all are compound nouns.

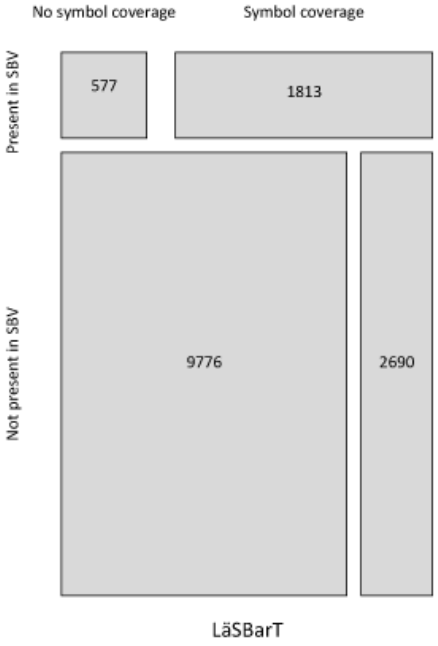


Figure 2: Overall ratio of LäSBarT nouns, presence in SBV and symbol coverage

### 5.3 Adjectives and adverbs

For adjectives, the proportion of lemmas without symbol coverage was as high as 29 %, while 40 % of the adverbs lacked symbol support. Differences in part-of-speech tagging for the two corpora, at the procedural as well as the annotational level, might however have influenced these results. Verb participles are for instance often subject to inconsistent tagging.

## 6 Augmenting the AAC lexicon

The next step was to investigate to what extent SALDO could be of assistance when augmenting

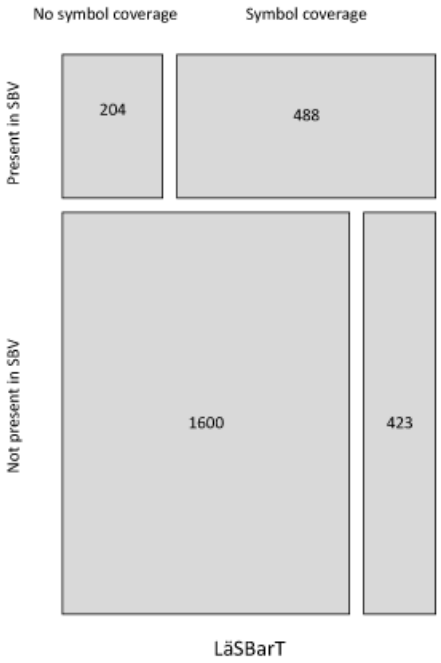


Figure 3: Overall ratio of LäSBarT adjectives, presence in SBV and symbol coverage

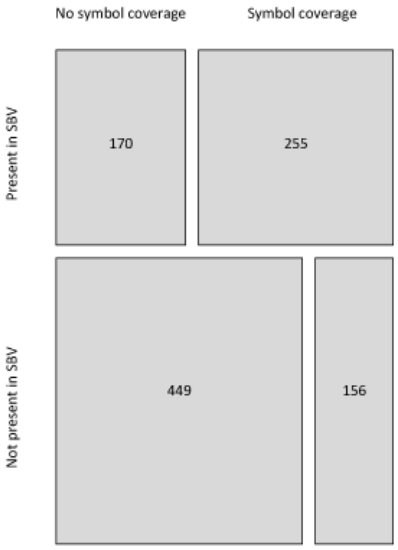


Figure 4: Overall ratio of LäSBarT adverbs, presence in SBV and symbol coverage

the AAC lexicon with additional concepts. Another interesting question concerned the feasibility of decomposing compounds into simplex words, each analysed against SALDO and provided with information necessary for correct symbol representation. Each entry in the set of lemmas present in both LBL and SBV, but without representation in the symbol lexicon, was compared against SALDO. As the concepts in SALDO are related by the mother-child relation, we could get the necessary lexical-semantic associations for further analysis of probable candidates for symbol representation. These could be either existing symbols, related as hyperonyms or synonyms, or a combination of two or more concepts.

As was stated earlier, a rather high proportion of noun lemmas missing in the symbol database were characterized as abstract nouns. We have for instance the noun lemma *kapitel* 'chapter', which had an adjusted frequency of 105.71 in SBV and a relative frequency of $1.03 \times 10^{-4}$ in LBL. From our core vocabulary database we get that the only existing entry is identified as *kapitel 1/1*, i.e. lemma identifier 1 and lexeme identifier 1. The next step is to consult SALDO, where a look-up of *kapitel* gives two matches: *kapitel..1* with the semantic descriptors *avsnitt* + *bok* 'section + book', and *kapitel..2*, with the semantic descriptor *kyrka* 'church'. Given the fact that in the primary corpus material, the word is unambigous, we allowed to illustrate the concept just by combining the symbols for *avsnitt* 'section' and *bok* 'book', both existing in the AAC database.

Concerning ompound nouns, which made up the largest portion of lemmas occurring only in LBL and not in SBV, (66 % of the 14,856 noun lemmas), decomposition into simplex words made it possible to achieve information enough for further elaboration into symbol representations. An example, illustrating this procedure, is the word *huvudkontor* 'head office'. It is not present in the symbol vocabulary, but we find it directly by a look-up in SALDO with the semantic descriptors *kontor* 'office' and *främst* 'major', both with symbol coverage in the database.

The last example is another compound noun, *affärsägare* 'shop owner', a word that does not exist in SALDO. The compound analysis tells that this word has two constituents with a linking morpheme,

namely *affär+s+ägare*. Since we already have the symbol illustrating the most common concept for *affär* in the primary corpus material, we use that. There is, however, no symbol in the database for *ägare*. Turning to SALDO,the word *ägare* 'owner' has only one descriptor *äga* 'to own'. We are now able to illustrate this concept by two symbols in combination, namely *affär* and *äga*, which by further analysis could possibly be extended to *person* 'person' + *äga* 'to own' + *affär* 'shop'.

As mentioned earlier, the few verbs not existing in the symbol database were generally either hapax, or particle verbs. Even if we regard the hapax words in LBL as peripheral in the easy-to-read genre, the fact that they exist in the SBV make them candidates for further analysis and inclusion into an augmented symbol lexicon. For nouns, the situation is largely the same. In general, they have a higher relative frequency, in average $8.0 \times 10^{-6}$, and only one third of the total are hapax words. Adjectives and adverbs in this set of words have a mean relative frequency in *LäSBarT* of $1.0 \times 10^{-5}$ and $4.4 \times 10^{-5}$, respectively. For adjectives, the hapax ratio was 30 % and for adverbs 20 %.

## 7 Conclusions

We found this to be a good way to produce a core vocabulary for Swedish. The suitability of this method was ensured not only by the fact that the ingoing entries were to be found in a corpus of simple texts, but also that they had a high degree of frequency and dispersion in a corpus balanced for genre and domain. It also turned out the the symbol coverage of these entries in the AAC language studied was impressively high for verbs (95 %), lower for nouns (76 %) and adjectives (71 %), and considerably lower for adverbs (60 %). This is completely in accordance with what we expected, since the basic verbs play a major role in communication. The fact that the nouns to a higher degree lack symbol support, was compensated for by the circumstance that a relatively high amount of entries could be found in or derived by information in a semantic lexicon.

Given that the results in this study are based on only one of several symbol languages, we would like to extend the research also to these, at first hand Bliss and more of the pictorial systems, such as PCS.

## References


Susan Balandin and T. Iacono. 1998. A few well-chosen words. *Augmentative and Alternative Communication*, 14:147–161.

Meher Banajee, Cynthia Dicarlo, and Sarintha Buras Stricklin. 2003. Core vocabulary determination for toddlers. *Augmentative and Alternative Communication*, 19(2):67–73.

D Beukelman and P Mirenda. 2005. *Augmentative and alternative communication: Supporting children and adults with complex communication needs.* Paul H. Brookes, Baltimore, 3rd edition.

Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of saldo and wordnet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, Odense.

Christiane Fellbaum. 1998. A semantic network of english - the mother of all wordnets. *Computers and the Humanities*, 32:209–220.

Eva Forsbom. 2006. A swedish base vocabulary pool. In *Swedish Language Technology conference*, Gothenburg.

W. Nelson Francis and Henry Kucera. 1979. Manual of information to accompany a standard corpus of present-day edited american english for use with digital computers. Technical report, Department of Linguistics, Brown University.

Dee Gardner. 2007. Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, 28(2):241–265.

David Hirsch and Paul Nation. 1992. What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, 8(2):689–696.

Stig Johansson, G. Leech, and H. Goodluck. 1978. Manual of information to accompany the lancaster-oslo/bergen corpus of british english, for use with digital computers. Technical report, Unversity of Oslo.

Katarina Mühlenbock. 2008. Readable, legible or plain words - presentation of an easy-to-read swedish corpus. In Anju Saxena and Åke Viberg, editors, *Multilingualism, Proceedings of the 23rd Scandinavian Conference of Linguistics*, Studia Linguistica Upsaliensia, ISSN 1652-1336; 8, pages 325–327, Uppsala, Sweden. Acta Universitatis Upsaliensis.

Inger Rosengren. 1972. *Ein Frekvenzwörterbuch der deutschen Zeitungssprache*. GWK Gleerup, Lund.

George Spache. March, 1953. A new readability formula for primary-grade reading materials. *Elementary School Journal*, LIII:410–413.

TEI Consortium. 2007. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* TEI Consortium, p5 edition.

Edward L. Thorndike and I. Lorge. 1944. *The teacher's word book of 30,000 words*. Columbia University Press, New York.

Åke Viberg, K Lindmark, A Lindvall, and I Mellenius. 2002. The swedish wordnet project. In *Proceedings of Euralex 2002*, pages 407–412, Copenhagen University.

M. West. 1953. *A General Service List of English Words.* Longman, London.

Widgit Software. 2011. Widgit homepage. http://www.widgit.com.

World Wide Augmentative & Alternative Communication. 2004. Communication is not a privilege. Technical report.