A Swedish Base Vocabulary Pool

Eva Forsbom
evafo@stp.lingfil.uu.se
Dep. of Linguistics and Philology, Uppsala University
Graduate School of Language Technology

SLTC, Göteborg, October 27-28, 2006

1 Motivation

In many language technology applications, there is a need for a base vocabulary, i.e. a vocabulary that could be reused for most domains and text types. For many languages, there exist one or more frequency dictionaries that contain some kind of base vocabulary, often based on a combination of the word's frequency and dispersion in a corpus.

A major problem with base vocabularies in frequency dictionaries is that they are not publicly available in electronic format. Another problem is that although such base vocabularies can be useful for some applications, the corpora they are based on (usually news articles) might not be representative for other applications, and so the base vocabularies will be of less use. A general-purpose, reusable, base vocabulary should therefore be based on a corpus with multiple genres and domains, such as a balanced corpus.

A third problem is what to include in, or rather exclude from, the base vocabulary. For some applications, the base vocabulary should be large (e.g. in machine translation and morphology analysis), but for others it should be small (e.g. for stylometry), or variable-sized (e.g. in computer-aided language learning). To make the base vocabulary reusable for several applications, we propose a base vocabulary pool, i.e. a ranked and annotated base vocabulary where the ranking is stable across genres and domains, and from which various subsets could be extracted for various needs.

2 Creation of the base vocabulary pool

2.1 Corpus

In the presentation, we will describe the derivation of a Swedish base lemma vocabulary pool from the 1-million word balanced corpora Stockholm-Umeå Corpus (SUC, 2002). SUC is compiled in a manner similar in spirit to that of the Brown (Francis and Kučera, 1979) corpus, and is meant to be representative of what a person might read in a year in the early nineties. Each word in SUC is annotated with its baseform and

its part-of-speech (mapped to the PAROLE tagset). The texts are also categorised in 9 major categories (genres) and 48 subcategories (domains).

The units of the base vocabulary pool are "lemmas", or rather the baseforms from the SUC annotation disambiguated for part-of-speech, so that the preposition *om* 'about' becomes `om.S` and the subjunction *om* 'if' becomes `om.CS`. Lemmas are used in favour of wordforms to somewhat alleviate the problem of data sparseness during computation.

2.2 Ranking

The lemmas are ranked according to relative frequency weighted with dispersion, i.e. how evenly spread-out they are across the subdivisions of the corpus, so that more evenly-spread words with the same frequency are ranked higher. This is done to compensate for accidental peaks of frequency due to certain texts, domains or genres. The weighting scheme is based on the ideas, introduced by Juilland (e.g. in Juilland and Chang-Rodriguez (1964)), behind the base vocabularies of most frequency dictionaries of today, including the Swedish "Nusvensk frekvensordbok" (NFO, Allén (1971)).

Since we would like to be able to extract the top N lemmas and not get hundreds of lemmas having the same rank, we experimented with various weighting schemes to find the one that gave the most discriminatory ranking. We used 5 schemes for each of the 3 possible dispersion categories (genre, domain, and text):

- Frequency alone.
- Modified frequency (as defined in Juilland and Chang-Rodriguez (1964)).
- Adjusted frequency (or 'Korrigierte Frekvenz', as defined in Rosengren (1972)).
- Dispersion alone (the weighting factor in Modified frequency, based on standard deviation).
- Contribution alone (as defined in Allén (1971), i.e. number of categories the lemma occurs in).

Adjusted frequency (see equation 1) was found to give the most discriminatory ranking for SUC (see Table 1). Adjusted frequency has also been shown to work well for corpora with variable-sized pre-defined categories.

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

where

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

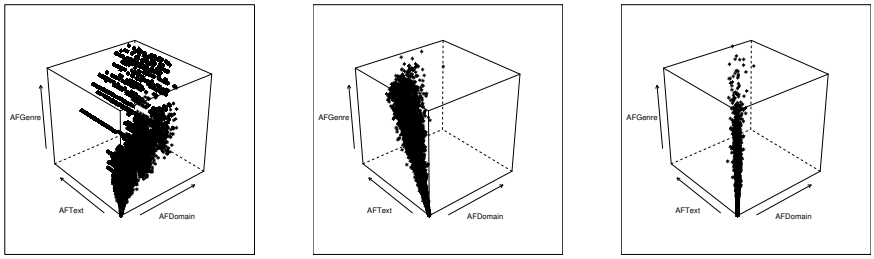
If we compare the ranking of the 100 top-ranked words with the ranking in NFO, the most notable difference, apart from some differences in classification, is the ranking of dialogue pronouns: *jag* 'I' (22 vs. 47 in NFO), and *du* 'you' (68 vs. 1200 in NFO).

Measure	Text	Domain	Genre
Adjusted frequency	18922	17044	10984
Modified frequency	3753	4136	4108
Dispersion	3452	3733	3781
Contribution	359	48	9
Frequency	572	572	572

Table 1: Number of ranks per measure for each category.

2.3 Filtering

The total vocabulary has 69,371 entries, but the base vocabulary pool is restricted to entries which occur in more than 3 genres, 8,215 entries. Although text and domain contribution filters were also tried, the filter on genre contribution gave the most visible differences. A threshold of 3 turned out to give the most stable ranking for adjusted frequency across three category divisions: genre, domain, text, while keeping as many lemmas as possible. As can be seen in Figure 1, there is a smooth 3-dimensional correlation cloud (i.e. no text- or domain-specific lemmas) for the 3-contribution filter, although it is not as thin as for the 8-contribution filter. For applications in need of smaller base vocabularies, a more restrictive filter could be used. The filter is corpus-specific, and depends on the number and characteristics of categories. In NFO, for example, lemmas occurring in more than 1 category were included in the base vocabulary. A 1-genre contribution filter also worked well in our experiments with the Susanne corpus (Sampson, 1995).



(a) No filter (69,371/100%)

(b) Contribution $>$ 3 (8,215/88.2%)

(c) Contribution $>$ 8 (1,505/74.7%)

Figure 1: Adjusted frequency rank correlation for text, domain, and genre in SUC, filtered by genre contribution. (Vocabulary size/Text coverage in parenthesis.)

The full base vocabulary pool covers 88.2% of the text in SUC, which is comparable to figures reported in Davies (2005): the frequency dictionary data from a 20-million word corpus of both written and spoken Spanish indicate that learners knowing about 4,000 lemmas will be able to recognize about 90% of all tokens in spoken Spanish (7,000 for 90% coverage in fiction, and 8,000 lemmas for non-fiction).

3 Applications and availability

We will use the base vocabulary pool as a basis for various subtasks in an information retrieval system. It has already been used for experiments in genre classification (Forsbom, 2005, 2006b), lemmatisation (Forsbom, 2006c), and morphological classification (Argaw and Forsbom, 2005), where it has shown to be useful.

The base vocabulary pool, including scripts for its creation and more details (Forsbom, 2006a), is available at `http://stp.lingfil.uu.se/~evafo/resources/basevocpool/`.

References

Sture Allén. *Nusvensk frekvensordbok baserad på tidningstext 2. Lemman. [Frequency dictionary of present-day Swedish based on newspaper material 2. Lemmas.]*. Data linguistica 4. Almqvist & Wiksell international, Stockholm, 1971.

Atelach Alemu Argaw and Eva Forsbom. Morphological classification of Swedish words using memory-based learning, 2005. Term paper for the GSLT course Machine Learning.

Mark Davies. Vocabulary range and text coverage. Insights from the forthcoming *Routledge Frequency Dictionary of Spanish*. In David Eddington, editor, *Selected Proceedings of the 7th Hispanic Linguistics Symposium*, pages 106–115, Somerville, MA, USA, 2005. Cascadilla Proceedings Project.

Eva Forsbom. Feature extraction for genre classification, 2005. Term paper for the GSLT course Statistical Methods.

Eva Forsbom. Deriving a base vocabulary pool from the Stockholm-Umeå corpus, 2006a. Term paper for the NGSLT course Soft Computing.

Eva Forsbom. Feature combination for genre classification, 2006b. Term paper for the course Artificial Neural Networks.

Eva Forsbom. Inducing baseform models from a Swedish vocabulary pool, 2006c. Term paper for the NGSLT course Minimally Supervised Morphology Induction.

W. Nelson Francis and Henry Kučera. *Manual of information to accompany a Standard Sample of Present-day Edited American English, for use with digital computers*. Providence, R.I., USA, 1979. Original ed. 1964, revised 1971, revised and augmented 1979.

Alphonse Juilland and E. Chang-Rodriguez. *Frequency Dictionary of Spanish words*. The Romance Languages and Their Structure, First Series S 1. Mouton & Co, The Hague, The Netherlands, 1964.

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

Geoffrey Sampson. *English for the Computer: The SUSANNE Corpus and analytic scheme*. Clarendon Press, Oxford, 1995.

SUC. Stockholm-Umeå corpus. Version 2.0. Stockholm University, Department of Linguistics and Umeå University, Department of Linguistics, 2002.