# Size is not Everything
# Genre Balance in Bootstrapping a Swedish PoS Tagger

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

## 1 Introduction

Part-of-speech tagging is a basic component of natural language processing, and as such, needs to be as accurate as possible, or any subsequent processing will suffer. For Swedish, most tagger models are trained on the Stockholm-Umeå Corpus (SUC Ejerhed et al., 2006). As SUC is a balanced corpus, SUC models are better representatives for general language than models trained on news texts only, which is a common scenario for other languages. On the other hand, the corpus is a bit too small for tagger training, considering the size of the tagset needed to express the most common morphosyntactic features of Swedish. This leads to poorer performance than what has been reported for, for example, an equally-sized English news corpus and a much smaller German news corpus, tagged with the statistical TnT tagger (Brants, 2000). Both the English and German models show an accuracy of 96.7%, while the same tagger trained on SUC only has an accuracy of 95.5%.

As SUC obviously is too small to be used alone as training data for any higher-accuracy tagger, we have used it as a seed corpus to bootstrap a much larger, unannotated, corpus, that can be added as training data. The bootstrapped corpus could represent another modality, domain or genre, if we are looking for adaptation. This sort of bootstrapping process has proved to be a viable approach (cf. Forsbom, 2006; Merialdo, 1994; Nivre and Grönqvist, 2001; Sjöbergh, 2003). Here, we are interested in seeing the effect the genre balance of bootstrapped corpus has on the performance, when drilling down by SUC genres.

## 2 Experimental Setup

The following bootstrap procedure was used:

1. Train a training model on all SUC.
2. Tag the bootstrap corpus using the training model.
3. Train an evaluation model on the tagged bootstrap corpus (not including SUC). For other taggers than TnT, train a TnT lexical model on the same data, to use for evaluation statistics on known/unknown words.
4. Evaluate the evaluation model on 10 folds of SUC, drilled-down by genre[1].
5. (Train a final tag model on a concatenation of all SUC and the tagged bootstrap corpus.)

For training and tagging, we used TnT and the new open-source implementation HunPos (Halácsy et al., 2007), with standard settings.

The labelled SUC corpus is a balanced corpus of modern Swedish prose covering approximately 1.2 million word tokens. The 1,040 text samples are meant to mirror what a Swedish person might read in the early nineties.

The distribution of texts and tokens between genres is shown in Table 1.

| ID | Genre | Samples (%) | Tokens (%) |
|---|---|---|---|
| a | Press: Reportage | 25.9 | 9.1 |
| b | Press: Editorial | 6.7 | 3.5 |
| c | Press: Reviews | 12.2 | 5.6 |
| e | Skills and Hobbies | 11.9 | 11.5 |
| f | Popular Lore | 6.0 | 9.4 |
| g | Biographies, essays | 2.6 | 5.2 |
| h | Miscellaneous | 13.9 | 13.9 |
| j | Learned and scientific writing | 8.3 | 16.4 |
| k | Imaginative prose | 12.5 | 25.4 |

Table 1: Distribution of texts and tokens per genre in SUC.

[1] URL http://stp.lingfil.uu.se/~evafo/software/cross_validation_sets

We tried the following large enough available corpora for bootstrapping:

- Europarl (E), version 2 (Koehn, 2005)[2]. It is a corpus of parallel texts in 11 languages from the proceedings of the European Parliament, 1996–2003. We have used the Swedish part (about 23.5 million words).
- Parole (P)[3]. It is a balanced Swedish corpus of about 19 million words (of which 4.4 are from novels, 13.6 from news papers, 0.4 from popular science, and 1 from web texts).
- Scarrie (S). It is a Swedish news text corpus of about 77 million words (Dahlqvist, 1999; Ohlander, 2005). The texts are from two daily papers: Svenska Dagbladet and Upsala Nya Tidning, 1995–1996.

## 3 Results

In Figure 1, overall an individual genre results are shown. As can be seen, all bootstrapped corpora are better than the baseline SUC, particularly for genres where SUC have little data.

Europarl, almost equal in size to Parole, but including only one genre not present in SUC, does fairly good on the c, g, and j genres, while it does even worse than baseline SUC for the k genre. Scarrie, on the other hand, which is almost three times larger than Parole, but only includes press genres, is not much better than Parole, except for the g and j genres. Combined corpora do not improve much over the best corpus included.

TnT is somewhat better for baseline SUC, while HunPos is better for the bootstrapped corpora, as HunPos also takes lexical probability into account.

In conclusion, Parole, or a similarly balanced corpus of equal or larger size, seems to be the best choice if we wish to represent general language.

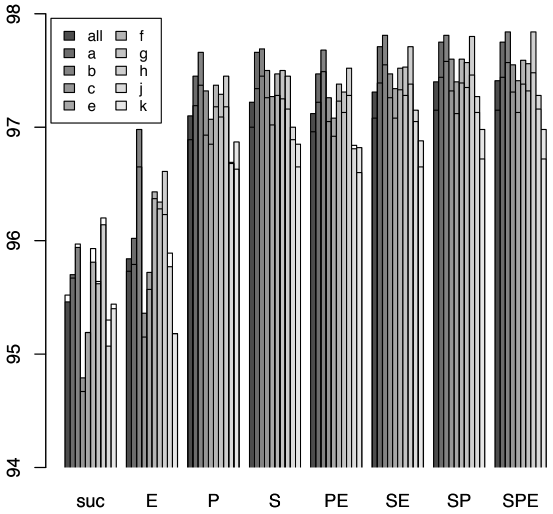

Figure 1: Estimated overall accuracy by SUC genre: baseline SUC, bootstrapped models for single or combined corpora. TnT results are shown on top of HunPos results.

## References

T. Brants. TnT - a statistical part-of-speech tagger. In *Proceedings of ANLP-2000*, Seattle, Washington, 2000.

B. Dahlqvist. A Swedish text corpus for generating dictionaries. In *The SCARRIE Swedish Newspaper Corpus*. Dep. of Linguistics, Uppsala University, 1999.

E. Ejerhed, G. Källgren, and B. Brodda. Stockholm-Umeå corpus version 2.0. Stockholm University, Dep. of Linguistics and Umeå University, Dep. of Linguistics, 2006.

E. Forsbom. Big is beautiful: Bootstrapping a PoS tagger for Swedish, 2006. Poster presentation at GSLT retreat, Gullmarsstrand.

P. Halácsy, A. Kornai, and C. Oravecz. Hunpos – an open source trigram tagger. In *Proceedings of ACL'07*, Prague, Czech Republic, 2007.

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit 2005*, Phuket, Thailand, 2005.

B. Merialdo. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2), 1994.

J. Nivre and L. Grönqvist. Tagging a corpus of spoken Swedish. *International Journal of Corpus Linguistics*, 6(1), 2001.

K. Ohlander. Lingvistisk annotering av tidningstext och extraktion av verbala konstruktionsegenskaper. Master's thesis, Dep. of Linguistics and Philology, Uppsala University, 2005.

J. Sjöbergh. Bootstrapping a free part-of-speech lexicon using a proprietary corpus. In *Proceedings of ICON-2003*, Mysore, India, 2003.

[2] URL: `http://www.statmt.org/europarl/`.

[3] The PAROLE corpus at The Swedish Language Bank, Göteborg University. URL: `http://spraakbanken.gu.se/parole/`.