

# Tagging a Corpus of Spoken Swedish

Joakim Nivre      Leif Grönqvist

Göteborg University  
Department of Linguistics  
P. O. Box 200  
405 30 Göteborg  
Sweden

## Abstract

In this article, we present and evaluate a method for training a statistical part-of-speech tagger on data from written language and then adapting it to the requirements of tagging a corpus of transcribed spoken language, in our case spoken Swedish. This is currently a significant problem for many research groups working with spoken language, since the availability of tagged training data from spoken language is still very limited for most languages. The overall accuracy of the tagger developed for spoken Swedish is quite respectable, varying from 95% to 97% depending on the tagset used. In conclusion, we argue that the method presented here gives good tagging accuracy with relatively little effort.

**Keywords:** statistical part-of-speech tagging, spoken language corpora

## 1 Introduction

Even though the number of corpora annotated for parts-of-speech has increased enormously over the past decade or so, most of these resources are still written language corpora. Consequently, researchers dealing with spoken language often find themselves in the situation of having to tag a corpus with no tagged corpus available to train the tagger. This was the problem we had to face in trying to tag the Gothenburg Spoken Language Corpus (GSLC), since no corpus of transcribed spoken Swedish had previously been tagged for parts-of-speech. Essentially, there seemed to be two alternatives available to us:

1. Tag a substantial part of the corpus manually and use that as training data for tagging the rest.
2. Use a training method that does not require tagged data as input, e. g., expectation-maximization using the Baum-Welch algorithm for Hidden Markov Models (Baum 1972).

None of these alternatives was very attractive, though. Manual tagging is a very time-consuming task and a lot of data is required to guarantee a good result; unsupervised learning does not require manual work but is known to

give poorer results in the end (Merialdo 1994). Therefore, we decided to try a different approach, namely to start by training a statistical tagger on written Swedish, for which several tagged corpora are now available, and to use various methods to adapt the resulting tagger to spoken language. Our belief in this strategy was strengthened by the results reported in Nivre *et al* (1996), where we used a tagger trained on written Swedish to tag spoken Swedish with no adaptations at all, and obtained encouraging if not brilliant results. Since then, we have refined the method to the point where we now actually obtain higher accuracy for spoken language than for written language. The purpose of this article is to present and evaluate our tagging methodology. Even though most of the work reported here concerns the analysis of a single corpus (GSLC), we believe that the results may be of more general interest, since the analysis of spoken language corpora is a rapidly growing research area and many research groups will have to face problems similar to ours in the future.

## 2 Background

### 2.1 The Gothenburg Spoken Language Corpus

The Gothenburg Spoken Language Corpus (GSLC) is a corpus of spoken Swedish collected at the Department of Linguistics, Göteborg University, over a number of years under the direction of Jens Allwood (cf. Allwood 1999a).<sup>1</sup> The corpus consists of audio and video recordings from a wide variety of social activities together with transcriptions of these recordings. At the time of writing, efforts are being made to digitalize the recordings and to synchronize recordings and transcriptions in order to create a multimodal spoken language corpus (cf. Nivre *et al* 1998). In this article we will restrict our attention to the corpus of transcriptions, which currently comprises some 1.2 million running words.

#### 2.1.1 Transcription Standard

Transcriptions in GSLC conform to the standard defined in Nivre (1999a), which contains a large number of symbols and conventions for marking, *inter alia*, speaker change, overlapping speech, pauses, and aspects of nonverbal communication through a system of standardized comments. Most of these features will not be of interest here, since they do not influence the assignment of parts-of-speech to word tokens, but we need to say a few words about the conventions for transcribing words.

Utterances are transcribed using *modified standard orthography* according to the following principles (cf. Nivre 1999b):

- Words that only occur with one pronunciation are always transcribed using standard orthography (but without capitalization). Thus, the words *Sverige* (Sweden) and *kanske* (maybe) are always transcribed **sverige** and **kanske**.
- Words that have several pronunciation variants are transcribed according to the actual pronunciation as far as this can be captured by modifying

---

<sup>1</sup>See also <http://www.ling.gu.se/SLSA/SLcorpus.html>.

the standard orthography. For example, the word *jag* (I) is transcribed as **ja** or **jag**, depending on whether the *g* is pronounced or not.

- In cases where the modified standard orthography gives rise to ambiguities (homographies) that are not present in standard orthography, the forms in question are disambiguated by one of two methods. Wherever possible, forms are disambiguated by supplying missing letters — with respect to standard orthography — in curly braces. Thus, the shorter form of *jag* is written **ja{g}** in order to distinguish it from the word *ja* (yes). In cases where this method of disambiguation is not applicable, numerical indices are used instead. For instance, the conjunction *och* (and) and the infinitive marker *att* (to) often have the same pronunciation, which is rendered **â0** (och) and **â1** (att).

These conventions enable us to capture the most important variations in pronunciation without having to use phonetic transcription. At the same time, we can in most cases convert the transcribed forms to standard orthography, which is often useful (e. g., in part-of-speech tagging).<sup>2</sup>

In addition to the conventions of modified orthography and disambiguation outlined above, the following options should be noted:

- Emphatic (or contrastive) stress is indicated by transcribing a word in all uppercase (e. g., **JAG** instead of **jag**).
- Lengthening of continuants is indicated by inserting a colon (:) after the corresponding grapheme (e. g., **ja:g**). Note that we refer to the extra lengthening here, not ordinary long vowels.
- Word forms that need to be disambiguated according to the principles outlined earlier, but where the transcriber is not sure what the intended word is are indexed with an asterisk \* (e. g., **â\***).
- Interrupted words are marked with a plus sign (+) at the end in order to distinguish them from (complete) homographs (e. g., **kon+**; cf. **kon** = ‘cone’).

### 2.1.2 Tokenization and Segmentation

Since the transcription system is based on standard orthography, segmentation of utterances into word tokens is performed in the process of transcription and word boundaries are marked by spaces as usual. This means that the problem of *tokenization*, i. e., the problem of determining what should count as one token and should therefore receive one part-of-speech tag, does not arise in the same way as for written texts where it is often unclear how, e. g., abbreviations should be split (or merged) into tokens. Of course, this does not mean that we avoid the problem that some fixed phrases, such as *på grund av* (because of), are perhaps better treated as single words (in this case as a preposition). As soon

---

<sup>2</sup>The only cases where it is not possible to convert transcribed forms to standard orthography are those where the modification does not create any ambiguity. For example, when an adjective like *roligt* (fun) is transcribed **rolit**, there is nothing in the form itself that tells us what the corresponding standard form is. In the future, we will try to avoid this problem by making curly braces obligatory in these cases as well: **rolit{g}t**.

as an expression is written as several words in standard orthography, it will be treated as such in our transcriptions as well.

Another property that distinguishes transcriptions from ordinary texts is the fact that they are clearly segmented into distinct utterances (or contributions), each of which belong to a certain speaker (even though some of these utterances may overlap in time). Since grammatical dependencies across utterances are normally rather weak (at least at the level of parts-of-speech), utterances provide a convenient unit for part-of-speech tagging (and utterance boundaries are an important contextual feature). Again, we can compare this to the task of segmenting a written text into sentences, which is clearly a non-trivial problem. (On the other hand, it is normally not necessary to do this segmentation in order to perform part-of-speech tagging.)

## 2.2 Statistical Part-of-Speech Tagging

Part-of-speech tagging refers to the problem of assigning lexical categories, or parts-of-speech, to words in a text, a problem for which there now exist a variety of methods. Sometimes, these methods are divided into two groups: statistical vs. rule-based. However, this terminology is somewhat misleading since many ‘rule-based’ methods rely on statistics in the training phase (e. g., Brill 1992) and some of the ‘non-statistical’ methods are not really rule-based in the traditional sense (e. g., Daelemans *et al* 1996). In this article, we use the term *statistical part-of-speech tagging* in a narrow sense, referring only to tagging methods that use a probabilistic model during the actual tagging phase and try to find the most probable part-of-speech sequence for a particular string of words.

### 2.2.1 HMM Tagging

Most statistical taggers are based on some variant of the  $n$ -class model (cf. Merialdo 1994), which can be seen as an instance of Shannon’s noisy channel model based on Bayesian inversion:

$$P(c_1, \dots, c_k | w_1, \dots, w_k) = \frac{P(w_1, \dots, w_k | c_1, \dots, c_k) P(c_1, \dots, c_k)}{P(w_1, \dots, w_k)}$$

In order to find the maximally probable part-of-speech sequence  $c_1, \dots, c_k$  for a given string of words  $w_1, \dots, w_k$ , we only need to find that sequence which maximizes the product in the numerator of the right hand side (since the denominator is constant for a given word string). The first factor of this product is given by the *lexical model*:

$$\hat{P}(w_1, \dots, w_k | c_1, \dots, c_k) = \prod_{i=1}^k P(w_i | c_i)$$

In this model, every word is conditioned only on its own part-of-speech, an independence assumption which may seem unrealistic but which is necessary in order to get a tractable and trainable model. Some early systems (e. g., DeRose 1988) instead use the inverse probabilities, i. e.,  $P(c_i | w_i)$ , which may be easier to estimate intuitively but which are not warranted by the noisy channel model and which appear to give worse performance (Charniak *et al* 1993).

The second factor is estimated by means of the *contextual model*:

$$\hat{P}(c_1, \dots, c_k) = \prod_{i=1}^k P(c_i | c_{i-(n-1)}, \dots, c_{i-1})$$

In this model, every part-of-speech is conditioned on the  $n - 1$  previous parts of speech. Depending on the value of  $n$ , we get different varieties of the  $n$ -class model, known as uniclass, biclass, triclass, etc. Earlier experiments have shown that while the triclass model generally yields better results given enough training data, the biclass model is more robust with respect to sparse data and therefore preferable for large tagsets and/or small training corpora (Nivre forthcoming).

The  $n$ -class model can be implemented very efficiently as a Hidden Markov Model (HMM), where the contextual model is defined by the transition probabilities of the underlying Markov chain, while the lexical model is defined by the output probabilities. The task of finding the most probable part-of-speech sequence for a given string of words is then equivalent to finding the optimal path (state sequence) of the model for a particular output string, a problem which can be solved with reasonable efficiency using the Viterbi algorithm (Viterbi 1967).

Given enough training data, statistical taggers based on the  $n$ -class model typically achieve accuracy rates ranging from 95% (Charniak *et al* 1993) to 97% (Merialdo 1994), depending on the type of text and the tagset used. Although most of the studies still concern English text, there is now a fair amount of studies reporting similar results for other languages, such as French (Chanod and Tapanainen 1995) and Swedish (Brants and Samuelsson 1995).

### 2.2.2 Parameter Estimation

The major problem in constructing a statistical tagger — or any other probabilistic model for that matter — is to find good estimates for the model parameters. In the  $n$ -class model, there are two types of parameters that need to be estimated:

1. Lexical probabilities:  $P(w|c)$
2. Contextual probabilities:  $P(c_i | c_{i-(n-1)}, \dots, c_{i-1})$

There are basically two methods that are used to estimate these parameters empirically from corpus data, depending on what kind of data is available for training. Both methods are based on the notion of Maximum Likelihood Estimation (MLE), which means that we try to choose those estimates that maximize the probability of the observed training data. If we have access to tagged training data, we can use relative frequencies to estimate probabilities:<sup>3</sup>

$$\hat{P}(w|c) = \frac{f_N(w, c)}{f_N(c)}$$

$$\hat{P}(c_i | c_{i-(n-1)}, \dots, c_{i-1}) = \frac{f_N(c_{i-(n-1)}, \dots, c_i)}{f_N(c_{i-(n-1)}, \dots, c_{i-1})}$$

---

<sup>3</sup>The relative frequency  $f_N(E)$  of an event  $E$  in a sample of  $N$  observations is always a maximum likelihood estimate of the probability  $P(E)$ ; see, e. g., Lindgren (1993).

If we only have access to untagged data, the standard method is to start from some initial model and use the Baum-Welch algorithm for Hidden Markov Models (Baum 1972) to iteratively improve the estimates until we reach a local maximum.<sup>4</sup> Unfortunately, there is no guarantee that we ever reach a *global* maximum, and results are generally better if we can use tagged data for estimation (Merialdo 1994).

Regardless of which method we use to obtain a maximum likelihood estimation from our training data, we still have to face the ubiquitous problem of *sparse data*, which means that, for a lot of the events whose probability we want to estimate, we simply do not have enough data to get a reliable estimate. The most drastic case of this is events that do not occur at all in the training data, such as ‘unknown words’ in the context of part-of-speech tagging. If we assign these events zero probability (according to MLE), then any chain of independent events involving such an event will also be assigned probability zero, which is usually not very practical (unless we can be sure that the event in question is really impossible and not just infrequent). Therefore, we normally want to adjust our estimates in such a way that we can reserve some of the probability mass for events that we have not yet seen. This is what is known in the field as *smoothing*.

### 2.2.3 Smoothing

Before we turn to the various methods used for smoothing, let us note that the problem of sparse data affects the two models involved in statistical part-of-speech tagging rather differently. In the contextual model, we always know how many events we have not seen. For example, given a part-of-speech system with  $N_C$  tags, we know that there are  $(N_C)^n$  possible  $n$ -tuples. By contrast, the lexical model is open-ended, and it is usually very difficult to estimate how many words (or word-tag pairs) we have not seen — unless we use a lexicon to stipulatively limit the class of words allowable in texts, a move which is often made when evaluating taggers, but which is usually completely unrealistic from a practical application point of view. We will return to the problem of the open-ended lexical model in section 3.

The methods used for smoothing can be divided into two broad categories. In the first category, which we may call *smoothing proper*, we find methods where the parameters of a single model are being adjusted to counter the effect of sparse data, usually by taking some probability mass from seen events and reserving it for unseen events. This category includes methods such as *additive smoothing* (Lidstone 1920, Gale and Church 1990), *Good-Turing estimation* (Good 1953, Gale and Sampson 1995), and various methods based on held-out data and cross-validation (Jelinek and Mercer 1985, Jelinek 1997).

In the second category, which we may call *combinatory smoothing*, we find methods for combining the estimates from several models. The most well-known methods in this category are probably *back-off smoothing* (Katz 1987) and *linear interpolation* (Brown *et al* 1992).

Most of the results on smoothing methods in the literature concern the problem of language modeling, i. e., of assigning probabilities to strings of words,

---

<sup>4</sup>The Baum-Welch algorithm can be seen as a special case of the general technique known as Expectation-Maximization (EM); cf. Dempster *et al* (1977).

a crucial problem in most statistical approaches to speech recognition (cf. Jelinek 1997, Ney *et al* 1997). For part-of-speech tagging, Nivre (forthcoming) has shown that the two models involved require different smoothing methods. While Good-Turing estimation seems to give the best results for the lexical model, this method is outperformed by both additive smoothing and back-off smoothing when it comes to the contextual model.

## 2.3 Previous Work on Swedish and Spoken Language

Although we have not been able to find any previous studies dealing specifically with spoken Swedish (except Nivre *et al* 1996), we should mention that there is a host of relevant work dealing with part-of-speech tagging either for (written) Swedish or for spoken languages other than Swedish (notably English).

Most important for our own work is the Stockholm-Umeå Corpus (Ejerhed *et al* 1992), a balanced corpus of written Swedish containing 1.2 million words. The corpus has been automatically tagged for parts-of-speech and manually corrected but is known to contain a small percentage of errors. The tagged version of the Stockholm-Umeå Corpus (SUC, for short) has been the primary source of training data for our lexical model (cf. section 3.5). Other studies dealing with part-of-speech tagging for Swedish include Samuelsson (1994), Brants and Samuelsson (1995), and Källgren (1996).

One of the pioneers in part-of-speech tagging of (transcribed) spoken language was Eeg-Olofsson (1991), who used statistical methods to tag a subpart of the London-Lund Corpus (Svartvik and Quirk 1980). In this study, however, the same (manually tagged) data was used in both training and testing, which means that the sparse data problem was not confronted head-on. Moreover, the task was facilitated by the fact that all the words to be tagged were in standard orthography. In more recent years, the biggest project in this area is without doubt the tagging of the spoken part of the British National Corpus, which was done using the CLAWS tagger (Garside 1987), a statistical tagger originally trained on the tagged version of the Brown Corpus. Some of the adaptations necessary in order to process spoken language with a written language tagger like the CLAWS tagger are described in Garside (1995), but on the whole there is not much information about this problem to be found in the literature.

# 3 Method

## 3.1 Training Data

The primary training data used to derive maximum likelihood estimates for lexical and contextual probabilities was the tagged version of the Stockholm-Umeå Corpus (cf. section 2.3), consisting of 1,166,902 tokens (punctuation included) and 107,075 distinct types. Before training the tagger, all words were converted to lowercase letters, which means that words that differed only in terms of capitalization were treated as occurrences of the same word type during training. This reduced the number of distinct types in the training corpus to 97,323.

Table 1 shows the number of tokens and types occurring with a specific number of different tags in the training corpus (using the basic SUC tagset of 23 tags; cf. section 3.2). On the one hand, we can see that 95.86% of all word

types only occur with one tag. On the other hand, these types together only represent 55.63% of the tokens in the corpus. The most ambiguous word type in the corpus is *i* (in), which occurs with 9 different tags, followed by *om* (about, if) and *för* (for) with 7 tags each.<sup>5</sup> The average number of tags per token is 2.04.

Table 1: Ambiguous types and tokens in the training corpus (large tagset)

Tags	Types	Tokens
1	93297	649205
2	2662	235902
3	291	131011
4	57	66693
5	8	7701
6	5	26614
7	2	21359
8	0	0
9	1	28417
Total	1166902	97323

### 3.2 Tagsets

The tagset used to tag the Stockholm-Umeå Corpus consists of 23 basic parts-of-speech with morpho-syntactic features that bring the total number of distinct tags to 156. In tagging the GSLC, we have not made any use of features, which means that the basic training tagset contains 23 distinct parts-of-speech. However, in tagging the spoken language corpus we have added two new parts-of-speech, *feedback* (fb) and *own communication management* (ocm). The class of feedback words includes primary feedback words, such as *ja* (yes) and *nej* (no), as well as secondary feedback words, e. g., adverbs like *precis* (precisely) when these are used to give feedback (cf. Allwood, Nivre and Ahlsén 1992). In the class of own communication management words we find mainly expressions used for hesitation and self-editing, such as *eh*, *öhh*, etc. (cf. Allwood, Nivre and Ahlsén 1990). We have also omitted two of the original SUC tags, *delimiter* (dl) and *foreign word* (uo). Delimiters (punctuation, etc.) simply do not occur in the spoken language corpus, while foreign words are to be tagged with their own part-of-speech (e. g., determiner [dt] for *the*). For the purpose of constructing a frequency dictionary of spoken Swedish (Allwood 1999b), we have also tagged the corpus with a smaller tagset of 11 parts-of-speech, corresponding to the traditional parts-of-speech found in grammar books, extended with fb and ocm. The two tagsets used for tagging the spoken language corpus are listed in appendix A and B.

<sup>5</sup>Interestingly, all three of these word types have pp (preposition) as their most frequent tag.

### 3.3 Tokenization and Segmentation

As mentioned earlier (cf. section 2.1.2), most of the tokenization in the spoken language corpus is done already during the transcription phase. However, the following points deserve to be noted:

- No attempt is made at “cleaning up” the transcriptions with respect to such phenomena as false starts, interrupted words, hesitation sounds, etc. All these phenomena, as long as they can be transcribed using modified standard orthography, are treated as words to be tagged.
- By contrast, all pauses and inaudible segments are filtered out from the utterances prior to tagging, and the resulting string of words is treated as a continuous utterance. This decision was based on the results in Nivre *et al* (1996), where the inclusion of pauses and inaudible segments did not improve tagging accuracy.

Transcriptions in GSLC are segmented into utterances (or contributions), where an utterance is defined as a continuous sequence of words (possibly including pauses), which is not interrupted by the non-overlapped speech of another participant.<sup>6</sup> During tagging, each utterance is fed to the tagger as a separate unit with no external context (except the start and end of the utterance) taken into account.

### 3.4 Tagger

The tagger used is a standard HMM triclass tagger using the Viterbi algorithm to derive the most probable part-of-speech sequence for a given string of words (utterance). We have chosen the triclass model over the biclass model in view of the results in Nivre (forthcoming), which indicate that with a tagset containing only 23 tags and a training corpus containing more than one million tokens, the triclass model is not seriously affected by the sparse data problem and performs significantly better than the biclass model with adequate smoothing.

In order to avoid underflow problems,<sup>7</sup> utterances are segmented into minimal strings with an unambiguous suffix of two words, i. e., where the last two words have only one possible part-of-speech each. By supplying these parts-of-speech as context for the next minimal string, we guarantee that no information is lost, since dependencies spanning more than three words are beyond the scope of the triclass model.

### 3.5 Lexical Model

#### 3.5.1 The SUC Model

The basic lexical model is given by the maximum likelihood estimates of  $P(w|c)$  derived from the training corpus and smoothed using Good-Turing estimation:

$$\hat{P}_{SUC}(w|c) = \frac{f_c^*(w)}{f(c)}$$

---

<sup>6</sup>In other words, the occurrence of overlap does not end the utterance of the first speaker until he/she stops speaking (cf. Nivre 1999a).

<sup>7</sup>Multiplying long chains of probabilities may yield numbers that are so small they are effectively rounded off to zero; this is known in the literature as the underflow problem (see, e. g., Cutting *et al* 1992).

In this equation,  $f(c)$  is the observed frequency of the part-of-speech  $c$  in the training corpus and  $f_c^*$  is the reestimation function for  $c$ , defined in the following way:

$$f_c^*(w) = (f_c(w) + 1) \frac{E(N_{f_c(w)+1})}{E(N_{f_c(w)})}$$

where  $f_c(w)$  is the observed frequency of word  $w$  with tag  $c$  in the training corpus, where  $N_{f_c(w)}$  and  $N_{f_c(w)+1}$  is the number of words occurring with tag  $c$  with frequency  $f_c(w)$  and  $f_c(w) + 1$ , respectively, and  $E(X)$  is the expectation value of the variable  $X$ . In practice, there is no way of precisely calculating expected frequencies of frequencies, and different versions of Good-Turing estimation differ mainly in the way they estimate these values from the observed frequencies of frequencies (see, e. g., Good 1953, Church and Gale 1991, Gale and Sampson 1995). In our case, the reestimated frequencies have been calculated using the simple Good-Turing method (Gale and Sampson 1995) in Dan Melamed’s implementation.<sup>8</sup>

We refer to the probability estimates derived in this way as the SUC lexical probabilities. In this model, unknown words (i. e., word forms not occurring in the training corpus) are treated as tokens of a single unknown word type  $w_u$  and assigned the following lexical probability for a given part-of-speech  $c$ :

$$\hat{P}_{SUC}(w_u|c) = \frac{1 - \sum_{w \in W} f_c^*(w)}{f(c)}$$

where  $W$  is the set of word forms occurring in the training corpus. Relying on the results in Nivre (forthcoming), the probability  $\hat{P}_{SUC}(w_u|c)$  was only defined for parts-of-speech  $c$  such that  $f(c) \geq 100$  and  $1 - \sum_{w \in W} f_c^*(w) \geq 0.001$ . (The parts-of-speech treated as open according to this definition are marked with an asterisk in appendix A and B.)

### 3.5.2 Handling Modified Standard Orthography

From the point of view of tagging the GSLC, the SUC lexical model in itself is deficient in that it does not take modified standard orthography into account, which means that forms like  $\mathbf{ja}\{\mathbf{g}\}$  and  $\mathbf{a0}$  will be treated as unknown words (i. e. as occurrences of  $w_u$ ). In order to remedy this, we need to define lexical probabilities for equivalence classes of word forms, where an equivalence class contains all the forms that correspond to a particular standard form. Let  $Std(w)$  denote the standard form corresponding to the (possibly modified) word form  $w$ .<sup>9</sup> As a first attempt, we can then define lexical probabilities as follows:

$$\hat{P}(w|c) = \hat{P}_{SUC}(Std(w)|c)$$

However, we still have to take into account the fact that transcribed words may be modified in order to capture aspects of prosody. In particular, it may be capitalized to signal heavy stress and/or contain colons that indicate lengthened segments. Let  $Pros(w)$  denote that word form which is exactly like  $w$  except that it is all in lowercase and contains no colon. (If  $w$  is already in lowercase

<sup>8</sup>Available at: [ftp://ftp.cis.upenn.edu/pub/melamed/tools/Good-Turing\\_smoothing/](ftp://ftp.cis.upenn.edu/pub/melamed/tools/Good-Turing_smoothing/).

<sup>9</sup>If the standard form of  $w$  is unknown (either because  $w$  is a genuinely unknown word or because it is an unknown modification of a known word), we let  $Std(w) = w_u$ .

and contains no colon, then obviously  $Pros(w) = w$ .) We can now define lexical probabilities in the following way:

$$\hat{P}(w|c) = \hat{P}_{SUC}(Std(Pros(w))|c)$$

### 3.5.3 Lexical Exceptions

Even when the problems of modified orthography and prosody have been taken care of, there is reason to believe that the SUC lexical model, which is based only on data from written Swedish, is in many respects inadequate for the analysis of spoken language. First and most importantly, for the two new parts-of-speech, *feedback* and *ocm*, we have no statistics at all from SUC. And even if we had, the estimates would probably not be very good, since these categories are much more frequent in spoken than in written language. Secondly, there are a number of specific word forms for which the SUC lexical probabilities are inadequate because they do not take all the relevant information into account. Let us consider three examples:

- In written Swedish, the word form *att* is ambiguous between the subordinate conjunction *att* (that) and the infinitive marker *att* (to). In spoken Swedish, the infinitive marker is very frequently realized as *å*, which in our transcriptions is disambiguated into *ã1*, whereas the subordinate conjunction is never realized in this way. This means that if we assign *ã1* the same lexical probabilities as written language *att* we will incorrectly predict that *ã1* is ambiguous between these two interpretations.
- Similarly, the word form *jag* is in written Swedish ambiguous between a first person pronoun (I) and a common noun (self). But the reduced form *ja{g}* occurring in spoken language is extremely improbable with a noun reading. Thus, we would like to say that the probability  $P(ja\{g\}|nn)$  is practically zero, although the probability  $P(jag|nn)$  is not.
- Finally, for personal pronouns in the third person plural, written language normally observes the case distinction between nominative *de* (they) and oblique *dem* (them), although the neutral form *dom* can also be found. In spoken language, except in certain dialects, the latter form is always used. Now, it so happens that this form is a homograph (but not a homophone) of the common noun *dom* (verdict). This ambiguity is as real in our transcriptions as in written language, but because of the much greater use of the forms *de* and *dem* in written language, the proportion between pronoun uses and noun uses for the form *dom* in written language is not at all representative for its use in spoken language.

What is common to all these cases is that we need to override the SUC lexical probabilities with manually constructed estimates. This is not so easy, however, since probabilities of the form  $P(w|c)$  are much more difficult to estimate intuitively than the “inverse” probabilities  $P(c|w)$ . Therefore, we rely on Bayesian inversion:

$$P(w|c) = \frac{P(c|w)P(w)}{P(c)}$$

If we can estimate  $P(c|w)$  manually, then we can compute  $P(w|c)$  if we know  $P(w)$  and  $P(c)$ . The class probability  $P(c)$  is usually not a problem (except for

**fb** and **ocm**), but the word probability  $P(w)$  may not be readily available if the frequency of  $w$  is radically different in written and spoken Swedish. On the other hand, when we are considering alternative parts-of-speech  $c$  for a given word  $w$  then  $P(w)$  is a constant, and as long as we are only interested in maximizing probabilities (rather than computing their exact values), we can omit  $P(w)$  from the earlier equation and rely on the following equation instead:

$$\hat{P}_{Exc}(w|c) = \frac{\hat{P}_{Man}(c|w)}{\hat{P}(c)}$$

(where  $\hat{P}_{Man}(c|w)$  is a manual estimate of  $P(c|w)$ ). However, this requires that we do not compare manually estimated probabilities with probabilities from the SUC lexical model, which means that if a word form requires a manually estimated probability for one part-of-speech  $c$ , then we must give the corresponding probabilities  $P(c'|w)$  for all permissible parts-of-speech  $c'$ . Thus, for the lexical exceptions discussed above, we currently use the following manually estimated probabilities:

- $\hat{P}_{Man}(ie|\hat{\mathbf{a}}1) = 1$
- $\hat{P}_{Man}(pn|ja\{\mathbf{g}\}) = 1$
- $\hat{P}_{Man}(pn|\mathbf{dom}) = 0.49995$
- $\hat{P}_{Man}(dt|\mathbf{dom}) = 0.49995$
- $\hat{P}_{Man}(nn|\mathbf{dom}) = 0.0001$

Altogether, we use manually estimated probabilities for some three hundred word forms. About eighty per cent of these are words having **fb** or **ocm** as one of their possible parts-of-speech. Since the Swedish lexical inventory of **fb** and **ocm** words has been studied in some detail (see, e. g., Allwood, Nivre and Ahlsén 1990, 1992), a list of the relevant items could be obtained rather easily.<sup>10</sup> The remaining group of sixty cases fall into three subgroups:

1. Lexical exceptions like  $\hat{\mathbf{a}}1$ ,  $ja\{\mathbf{g}\}$  and  $\mathbf{dom}$ , discussed earlier, where the statistical estimates derived from the written language corpus are especially problematic.
2. Foreign words, which in SUC have a special tag (**uo**), while we prefer to tag them with their actual part-of-speech (e. g., determiner [**dt**] for *the*).
3. Certain determiners, such as *många* (many) and *få* (few), which are classified as adjectives (**jj**) in SUC, while we prefer to view them as determiners (**dt**).

The words in the second group (lexical exceptions, foreign words, and determiners/adjectives) have been identified using the researchers' intuition, aided by an iterative process of tagging the spoken corpus and checking for "suspicious cases" in the resulting frequency lists for different parts-of-speech, e. g.,  $\mathbf{dom}$  (they/them/verdict) turning up as one of the most frequent nouns.

One question that arises in constructing the manually estimated lexical probabilities is how to treat prosodic modifications. It is clear that the presence of

---

<sup>10</sup>Many of these word forms do not occur in SUC at all, so some sort of exceptional treatment would have been necessary in any case.

a certain prosodic feature may increase the probability of a particular part-of-speech assignment. Thus, the word *m* may occur either as *fb* (acknowledgement) or as *ocm* (hesitation marker), but the presence of lengthening (*m:*) would normally make the *ocm* interpretation more likely. In the current version of the tagger, we have not exploited prosodic information in this way, but we may wish to do so in the future. Furthermore, although some of the manually estimated probabilities concern specific variants of word forms (such as *ja{g}* as opposed to *jag*), in most cases we want to assign the same estimate to all variants of a word form (e. g., *de{t}* and *det*). Therefore, we have to give a three-step definition of lexical probabilities for exceptional words:

1.  $\hat{P}(w|c) = \hat{P}_{Exc}(w|c)$  if defined;
2. else:  $\hat{P}(w|c) = \hat{P}_{Exc}(Pros(w)|c)$  if defined;
3. else:  $\hat{P}(w|c) = \hat{P}_{Exc}(Std(Pros(w))|c)$  (if defined).

The first case applies to word forms containing significant prosodic modifications (such as the hypothetical *m:* mentioned earlier); the second to specific variants with or without prosodic modification (e. g., *ja{g}* and *ja:g*), which have a lexical probability different from *jag* and *ja:g*; the third to words where all variants of a word form have the same (manually estimated) lexical probabilities (e. g., *de:{t}*, *de:t*, *de{t}* and *det*).

#### 3.5.4 Ambiguous Words

In order to make the lexical model complete, we must also handle “star forms” like *â\**, which are ambiguous between several (known) word forms. Since the different alternatives are mutually exclusive, we can estimate the probability of a particular part-of-speech assignment simply by summing over the alternatives:

$$\hat{P}(w^*|c) = \sum_{w_i \in *_w} \hat{P}(w_i|c)$$

where  $*_w$  is the set of possible disambiguations of  $w^*$ .

#### 3.5.5 Interrupted Words

When a word is interrupted it is usually impossible to say with certainty which word was intended. We have therefore decided to treat all interrupted words (ending in *+*) as instances of the category *ocm*. Strictly speaking, this is not correct, since it is the interruption itself, rather than the interrupted word, which is an OCM phenomenon, but it seems unreasonable to let the tagger guess which part-of-speech was intended when in most cases this is beyond the capacity of a human interpreter. The lexical model must therefore be extended accordingly:

1.  $\hat{P}(w+|ocm) = \hat{P}_{SUC}(w_u|ocm)$
2.  $\hat{P}(w+|c) = 0$  for all  $c \neq ocm$

The effect of this treatment is that all interrupted words have zero probability of occurring with any tag other than *ocm*, which means that they are unambiguously assigned to this category.

### 3.5.6 Unknown Words

According to the lexical model defined so far, all unknown words are treated as tokens of the unknown word type  $w_u$  and assigned the probability  $\hat{P}_{SUC}(w_u|c)$  for all open parts-of-speech  $c$ . In practice, we have implemented one further heuristic to narrow down the range of possible parts-of-speech for unknown words:

If  $w$  can be parsed as a numeral, then it is assigned the probability  $\hat{P}_{SUC}(w_u|\text{rg})$  (but probability zero for all other open parts-of-speech).

### 3.5.7 Summary

Putting all the pieces together, we may now define the complete lexical model used in tagging the spoken language transcriptions as follows:

1. For every word form  $w^*$  ending in an asterisk (\*):  

$$\hat{P}(w^*|c) = \sum_{w_i \in *w} \hat{P}(w_i|c)$$
2. For every word form  $w+$  ending in a plus (+):  

$$\hat{P}(w+|\text{ocm}) = \hat{P}_{SUC}(w_u|\text{ocm})$$

$$\hat{P}(w+|c) = 0 \text{ for all } c \neq \text{ocm}$$
3. For every other word form  $w$ ,  $\hat{P}(w|c) =$ 
  - (a)  $\hat{P}(w|c) = \hat{P}_{Exc}(w|c)$  if defined;
  - (b) else:  $\hat{P}(w|c) = \hat{P}_{Exc}(Pros(w)|c)$  if defined;
  - (c) else:  $\hat{P}(w|c) = \hat{P}_{Exc}(Std(Pros(w))|c)$  if defined;
  - (d) else:  $\hat{P}(w|c) = \hat{P}_{SUC}(Std(Pros(w))|c)$ .

where

1.  $\hat{P}_{Exc}(w|c) = \hat{P}_{Man}(c|w)/\hat{P}(c)$
2.  $\hat{P}_{SUC}(w|c)$  is defined only for
  - (a)  $c$  such that  $f_c(w) > 0$  if  $f(w) > 0$ ;
  - (b)  $c = \text{rg}$  if  $f(w) = 0$  and  $w$  can be parsed as a numeral;
  - (c)  $c$  such that  $f(c) \geq 100$  and  $1 - \sum_{w \in W} f_c^*(w) \geq 0.001$  otherwise.

## 3.6 Contextual Model

### 3.6.1 The SUC Model

The basic contextual model is given by the maximum likelihood estimates of  $P(c_i|c_{i-2}, c_{i-1})$  derived from the training corpus and smoothed using simple additive smoothing (with  $k = 0.5$ ):<sup>11</sup>

$$\hat{P}_{SUC}(c_{i-2}, c_{i-1}, c_i) = \frac{f_{SUC}(c_{i-2}, c_{i-1}, c_i) + 0.5}{(N_{SUC} - 2) + 0.5 \cdot 25^3}$$

<sup>11</sup>This way of smoothing the maximum likelihood estimates is sometimes referred to as Expected Likelihood Estimation (ELE); cf. Gale and Church (1990).

$$\hat{P}_{SUC}(c_{i-1}, c_i) = \frac{f_{SUC}(c_{i-1}, c_i) + 0.5}{(N_{SUC} - 1) + 0.5 \cdot 25^2}$$

$$\hat{P}_{SUC}(c_i | c_{i-2}, c_{i-1}) = \frac{\hat{P}_{SUC}(c_{i-2}, c_{i-1}, c_i)}{\hat{P}_{SUC}(c_{i-2}, c_{i-1})}$$

where  $f_{SUC}$  refers to frequencies in the training corpus and  $N_{SUC}$  is the number of word tokens in this corpus. Additive smoothing was chosen because it was one of the two top scoring methods for the contextual model in Nivre (forthcoming) and because it is much easier to implement than the second method, back-off smoothing, especially when one wishes to include categories that are not present in the training corpus, such as `fb` and `ocm` in our case. We refer to the estimates derived in this way as the SUC contextual probabilities.

### 3.6.2 The GSLC Model

It seems very likely that the contextual probabilities derived from the written training corpus will not in all respects be representative for the patterns found in spoken language. In particular, this concerns the new categories `fb` and `ocm` for which we really have no data at all. In an attempt to overcome this problem, we define a second contextual model based on the results of tagging the spoken language corpus with the SUC contextual model and the lexical model defined in section 3.5. We refer to this model as the GSLC contextual model:

$$\hat{P}_{GSLC}(c_{i-2}, c_{i-1}, c_i) = \frac{f_{GSLC}(c_{i-2}, c_{i-1}, c_i) + 0.5}{(N_{GSLC} - 2) + 0.5 \cdot 25^3}$$

$$\hat{P}_{GSLC}(c_{i-1}, c_i) = \frac{f_{GSLC}(c_{i-1}, c_i) + 0.5}{(N_{GSLC} - 1) + 0.5 \cdot 25^2}$$

$$\hat{P}_{GSLC}(c_i | c_{i-2}, c_{i-1}) = \frac{\hat{P}_{GSLC}(c_{i-2}, c_{i-1}, c_i)}{\hat{P}_{GSLC}(c_{i-2}, c_{i-1})}$$

In these equations,  $f_{GSLC}$  refers to frequencies in the tagged spoken language corpus, while  $N_{GSLC}$  is the total number of word tokens in this corpus. In computing these frequencies, each utterance boundary is treated as two subsequent tokens of a special boundary symbol (`$`), tagged with the dummy category `start`, which is also supplied as left and right context for each utterance during tagging. This means that the conditioning of contextual probabilities includes the start and end of utterances but does not span across utterances.

## 3.7 Evaluation

The standard way of evaluating part-of-speech taggers (assuming a condition of forced choice) is by computing their accuracy rate, i. e., the proportion of correct tags out of the total number of tagged tokens:

$$\text{Accuracy} = \frac{\#\text{Tokens correctly tagged}}{\#\text{Tokens}}$$

Accuracy rates will be given with a .95 binomial confidence interval, and differences between taggers have been tested for significance using McNemar’s test (Everitt 1977, Dietterich 1997).

In addition, we will compute recall and precision for individual parts-of-speech  $c$ , defined in the following way:

$$\text{Recall}(c) = \# \text{Tokens correctly tagged } c / \# \text{Tokens of } c$$

$$\text{Precision}(c) = \# \text{Tokens correctly tagged } c / \# \text{Tokens tagged } c$$

The evaluation is based on a random sample of 822 utterances from the spoken language corpus consisting of 10,003 word tokens, which was tagged manually by one of the authors. During the manual tagging process, 3 complete utterances (comprising 244 word tokens) and 2 utterance segments (9 word tokens) were excluded from the test corpus because they were in a language other than Swedish. By contrast, isolated foreign words and short phrases occurring as part of a Swedish utterance were retained and scored as correct only if they were tagged with their correct part-of-speech (e. g., determiner [dt] for *the*). The final test corpus thus consisted of 819 utterances and 9,750 word tokens. Table 2 shows the distribution of different utterance lengths (measured in words per utterance) in the test corpus. In cases where the human annotator was not

Table 2: Utterance length distribution in the test corpus ( $N = 819$ )

No. of words	Frequency ( $f$ )	Proportion ( $f/N$ )
1	209	0.26
2	52	0.06
3	57	0.07
4	46	0.06
5	51	0.06
6–10	153	0.19
11–20	133	0.16
21–	118	0.14
Total	819	1.00

able to choose a single tag, the tag assigned automatically was scored as correct if it was one of the tags considered possible by the human annotator. These cases represented about 2.6% of the word tokens with the large tagset and about 1.7% with the small tagset. If the tag assigned automatically was *not* one of those considered by the human annotator, one of the potentially correct tags was randomly chosen as the correct one when computing the recall for different parts of speech. In this way, we ensure that the maximum recall (as well as precision) is always 100%.

## 4 Results

### 4.1 Overall Accuracy

Tagging with the SUC contextual model resulted in an accuracy rate of 93.85% ( $\pm 0.48\%$ ) for the large tagset and 96.16% ( $\pm 0.39\%$ ) for the small tagset. The corresponding results for the GSLC contextual model were 95.30% ( $\pm 0.43\%$ ) (large tagset) and 97.44% ( $\pm 0.32\%$ ) (small tagset). The difference between the two models is significant at the .99 level for both tagsets (McNemar’s test).

## 4.2 Recall and Precision

The recall and precision for different parts-of-speech in the large tagset are given in table 3. For the majority of categories (12 out of 22<sup>12</sup>) the GSLC contextual

Table 3: Recall and precision (large tagset)

Tag	Tokens	Recall <sub>SUC</sub>	Recall <sub>GSLC</sub>	Precision <sub>SUC</sub>	Precision <sub>GSLC</sub>
ab	1455	90.03	93.06	95.55	96.85
dt	406	90.89	93.35	82.55	81.68
fb	635	98.58	99.84	93.71	99.37
ha	101	92.08	92.00	54.71	66.91
hd	10	100.00	90.00	100.00	100.00
hp	176	90.34	90.34	92.98	94.64
ie	100	95.00	95.00	81.20	79.17
in	27	88.89	96.30	96.00	100.00
jj	322	82.30	88.20	93.31	94.67
kn	509	91.16	94.30	97.27	95.62
nn	986	93.91	97.82	95.66	95.73
ocm	215	100.0	100.00	96.41	100.00
pc	57	78.95	77.19	50.00	80.00
pl	113	81.42	75.22	78.63	82.52
pm	92	83.70	86.96	87.50	89.89
pn	1623	95.26	95.00	98.41	99.04
pp	607	95.06	97.20	91.30	90.49
ps	36	100.00	100.00	100.00	100.00
rg	118	90.68	95.76	99.07	99.12
ro	16	93.75	93.75	71.43	78.95
sn	203	87.19	84.24	86.34	89.53
vb	1943	98.02	99.02	98.61	98.76
Total	9750	93.85	95.29	93.85	95.29

model gives better results than the SUC contextual model for both recall and precision. Four categories have marginally lower recall but this is compensated by a gain in precision. Conversely, for six categories precision drops while recall improves or stays the same.

Looking at the actual figures for the best contextual model (GSLC), we find that 13 out of 22 categories in the large tagset have both recall and precision above 90%. Of these categories, **fb**, **ocm**, **ps** and **vb** have especially high figures (about 99% or better for both recall and precision). Of the remaining categories, four (**dt**, **ha**, **ie**, **ro**) have considerably better recall than precision and thus appear to be overused by the tagger. Conversely, one category (**jj**) has precision but not recall above 90%. The remaining four (**pc**, **pl**, **pm**, **sn**) have both precision and recall below 90%.

Table 4 gives the corresponding figures for the small tagset. Here the GSLC contextual model improves recall and precision for virtually all categories. The only exception is **prep**, which has a small drop in precision (but a gain in recall).

<sup>12</sup>One of the 23 categories, *wh-possessives* (hs), was not represented at all in the test corpus.

Table 4: Recall and precision (small tagset)

Tag	Tokens	Recall <sub>SUC</sub>	Recall <sub>GSLC</sub>	Precision <sub>SUC</sub>	Precision <sub>GSLC</sub>
adj	378	83.60	87.56	84.49	93.24
adv	1669	94.25	95.45	94.87	97.13
fb	635	98.58	99.84	93.71	99.37
int	27	88.89	96.30	96.00	100.00
conj	812	94.58	96.31	96.12	96.19
noun	1078	94.71	98.24	96.59	96.66
num	133	92.48	96.24	94.62	96.24
ocm	215	100.00	100.00	96.41	100.00
pron	2253	98.09	98.45	98.93	99.28
prep	607	95.06	97.20	91.30	90.49
verb	1943	98.92	99.02	98.61	98.76
Total	9750	96.16	97.44	96.16	97.44

With the GSLC contextual model, both recall and precision are over 90% for all categories except one (`adj`).

## 5 Discussion

Let us begin by noting that the overall accuracy rates of our spoken language taggers compare fairly well with the performance reported in the literature for written language (cf. section 2.2.1), ranging from 94% to 97% depending on tagset and contextual model. Not surprisingly, we also see that the reestimation of the contextual model based on a preliminary tagging of the spoken language corpus yields a substantial and highly significant improvement of overall accuracy.

This result indicates not only that there are important differences between the syntax of written and spoken language — which is hardly surprising — but also that these differences are reflected even in a syntactic model as crude as the probabilistic triclass model. A good case in point is the preponderance of one-word utterances in spoken language (cf. section 3.7) with probability distributions differing greatly from the ones found in written language. An important conclusion of our work must therefore be that a reestimation of the contextual model is a crucial step in the adaptation process if we want to apply a statistical tagger trained on written language to the analysis of spoken language data.

Looking at individual parts-of-speech, we see that recall/precision is improved more or less across the board, but improvement is especially noticeable for the typical “spoken language categories” `fb` (precision), `in` (recall and precision), and `ocm` (precision). Conversely, categories that are more frequent in written than in spoken language often have worse precision with the GSLC contextual model. Cases in point are `dt`, `ie`, and `pp`. (With the small tagset, these tendencies disappear due to the fact that these categories are not clearly disambiguated as such; see below.)

From a methodological point of view, it is interesting to note that the im-

provement due to the retraining of the contextual model occurs despite the fact that the preliminary tagging contains a fair amount of tagging errors (4–6% depending on tagset). This raises the question whether it is possible to improve the results further by iterating this procedure, i. e., by retraining the contextual model again based on the results of the second tagging, etc. It appears that this is not the case. Preliminary experiments have shown that there are few significant changes in the contextual model as a result of further iterations and that performance, if anything, gets worse. This probably points to an overtraining effect.

Crucial as the contextual model may be, it is nevertheless the case that the quality of the lexical model affects tagging accuracy even more. Fortunately, our results seem to indicate that a lexical model derived from written language data can be adapted to spoken language with rather minimal efforts, i. e., by providing manual probability estimates for a few hundred exceptional words.

The results look even more promising when we consider the fact that we have so far done very little to cope with the ubiquitous problem of unknown words, which means that there should be room for further improvement by implementing standard heuristics such as looking at word endings (see, e.g., Samuelsson 1994).

Another way of improving tagging accuracy is to try to eliminate especially frequent errors. If we consider the recall/precision figures for individual parts-of-speech with the GSLC contextual model, we find that some tags are overused (higher recall than precision), while others are underused (higher precision than recall). To some extent, these categories come in pairs; and to some extent the error patterns for these pairs seem to reflect a “written language bias” in the lexical model for certain word forms. The pattern is especially clear with the large tagset where, for example, pronouns (pn) are frequently misclassified by the tagger as determiners (dt), ordinary adverbs (ab) as *wh*-adverbs (ha), verb particles (pl) and conjunctions (kn) as prepositions (pp), the subjunction *att* (sn) as an infinitive marker (ie), and the relative pronoun *som* (hp) as a conjunction (kn). In all these cases, the frequency distribution of different parts-of-speech for a given word form differs significantly between written and spoken Swedish and, since the lexical model is derived from written language data, the tagger tends to err in the direction of overusing the tags that are more frequent in written language than in spoken language, even with the reestimation of the contextual model.

Errors belonging to the five “confusion pairs” listed above represent more than 40% of all errors made by the tagger with the large tagset, 189 cases out of a total of 458 errors in the test corpus. Moreover, of these 189 errors, almost 80% (150) are contributed by as few as eight word forms. These word forms are listed, in order of decreasing error frequency, with rough English translations in table 5. The fact that such a small number of word forms account for such a large part of the tagging errors may be seen as a positive result in the sense that, in principle, it seems possible to improve the performance of the tagger considerably by fine-tuning the lexical model for a relatively small number of word forms. However, it should also be pointed out that, for all of the word forms listed in table 5, there is also a small number of errors going in the other direction.<sup>13</sup> It is therefore an open question to what extent it is possible, by

---

<sup>13</sup>The ratio between the two error types is 1:6 on average.

Table 5: Eight frequent error types

Word form	Correct tag	Erroneous tag	Frequency
det	pn: <i>it</i> <sub>NEUTER</sub>	dt: <i>the</i> <sub>NEUTER</sub>	33
då	ab: <i>then</i>	ha: <i>when</i> <sub>REL</sub>	28
att	sn: <i>that</i>	ie: <i>to</i>	23
där	ab: <i>there</i>	ha: <i>where</i> <sub>REL</sub>	18
som	hp: <i>who/which/that</i>	kn: <i>as</i>	16
den	pn: <i>it</i> <sub>UTER</sub>	dt: <i>the</i> <sub>UTER</sub>	13
för	kn: <i>for</i>	pp: <i>for</i>	11
dom	pn: <i>they/them</i>	dt: <i>the</i> <sub>PLUR</sub>	8

hand-tuning lexical probabilities, to reduce the one type of error without a corresponding increase in the other type. For the moment, we have no data on this, but it is clearly a problem that is worth exploring further.

Finally, concerning the small tagset, we simply observe that the generally higher recall/precision figures (as compared to the large tagset) are due to the fact that several of the difficult tag pairs discussed above are simply merged in the smaller tagset. Thus, determiners and pronouns are both classified as pronouns (*pron*),<sup>14</sup> ordinary adverbs and *wh*-adverbs as adverbs (*adv*), etc.

## 6 Conclusion

Many research groups working with spoken language have to face the problem of tagging a corpus of transcribed spoken language for parts-of-speech with no tagged corpus of spoken language available to train the tagger. In this article, we have tried to show that one way of solving this problem is to go through the following five steps:

1. Train a statistical tagger on tagged data from written language.
2. Adjust the lexical model in the following two ways:
  - (a) Define a mapping of non-standard word forms in the transcriptions (if any) to written standard forms.
  - (b) Define an exceptional lexical model for word forms where the probabilities derived from written language data can be expected to differ greatly from their true probabilities in spoken language.
3. Tag the spoken language corpus using the adjusted lexical model and the contextual model derived from written language data.
4. Retrain the contextual model on the tagged spoken language corpus resulting from step 3.
5. Tag the spoken language corpus using the adjusted lexical model and the newly derived contextual model.

<sup>14</sup>This is standard practice in Swedish elementary grammars, probably due to the large number of word forms that can be used in both ways.

The results of our evaluation indicate that this method can produce good-quality tagging and therefore represents a viable alternative to more traditional methods of supervised and unsupervised training of statistical taggers. Since there is no need to tag large amounts of training data manually, our method is less labor-intensive than ordinary supervised learning — in the context envisaged — while apparently giving comparable accuracy. And since unsupervised learning generally tends to give lower accuracy than supervised learning (Merialdo 1994), we can at least hypothesize that our method is preferable to, say, Baum-Welch training on grounds of accuracy.

To test this hypothesis, or indeed to make a comparative study of all three methods, is an interesting project for the future. In the meantime, we can only conclude that our method seems to offer a good compromise between the demands of minimizing effort and maximizing accuracy, and we can therefore recommend it to research groups facing problems similar to ours.

## References

- Allwood, J. 1999a. “The Swedish Spoken Language Corpus at Göteborg University.” In *Fonetik 99: Proceedings from the Twelfth Swedish Phonetics Conference*, Gothenburg Papers in Theoretical Linguistics 81. Göteborg University: Department of Linguistics. (Also available at: <http://www.ling.gu.se/fonetik99/manus/31.ps>.)
- Allwood, J. (ed) 1999b. “Talspråksfrekvenser. Ny och utvidgad upplaga.” Gothenburg Papers in Theoretical Linguistics S21. Göteborg University: Department of Linguistics.
- Allwood, J., Nivre, J. & Ahlsén, E. 1990. “Speech Management—On the Non-Written Life of Speech.” *Nordic Journal of Linguistics* 13, 3–48.
- Allwood, J., Nivre, J. & Ahlsén, E. 1992. “On the Semantics and Pragmatics of Linguistic Feedback.” *Journal of Semantics* 9, 1–26.
- Baum, L. E. 1972. “An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of a Markov Process.” *Inequalities* 3, 1–8.
- Brants, T. & Samuelsson, C. 1995. “Tagging the Teleman Corpus.” In *Proceedings of the 10th Nordic Conference of Computational Linguistics, NODALIDA-95*, Helsinki, 7–20.
- Brill, E. 1992. “A Simple Rule-based Part of Speech Tagger.” In *Third Conference of Applied Natural Language Processing*, ACL.
- Brown, P. F., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L. and Roossin, P. S. 1992. “Class-Based *N*-gram Models of Natural Language.” *Computational Linguistics* 18, 467–479.
- Chanod, J.-P. & Tapanainen, P. 1995. “Tagging French — Comparing a Statistical and a Constraint-based Method.” In *Seventh Conference of the European Chapter of the Association for Computational Linguistics*, Dublin, 149–156.
- Charniak, E., Hendrickson, C., Jacobson, N. and Perkowitz, M. 1993. “Equations for Part-of-Speech Tagging.” In *Proceedings of the Eleventh National Con-*

*ference on Artificial Intelligence*, AAAI Press/MIT Press.

Church, K. W. and Gale, W. A. 1991. "A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilities of English Bigrams." *Computer Speech and Language* 5, 19–54.

Cutting, D., Kupiec, J., Pedersen, J. and Sibun, P. 1992. "A Practical Part-of-speech Tagger." In *Third Conference on Applied Natural Language Processing*, ACL, 133–140.

Daelemans, W., Zavrel, J., Berck, P. and Gillis, S. 1996. "MBT: A Memory-Based Part of Speech Tagger-Generator." In Ejerhed, E. and Dagan, I. (eds) *Proceedings of the Fourth Workshop on Very Large Corpora*, Copenhagen, Denmark, 14-27.

Dempster, A. P., Laird, N. M. and Rubin, D. B. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society* 39, 1–38.

DeRose, S. J. 1988. "Grammatical Category Disambiguation by Statistical Optimization." *Computational Linguistics* 14, 31–39.

Dietterich, T. G., 1998. "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms." *Neural Computation*, 10 (7) 1895-1924.

Eeg-Olofsson, M. 1991. "Probabilistic Tagging of a Corpus of Spoken English." In Eeg-Olofsson, M. 1991. *Word-Class Tagging: Some Computational Tools*. University of Göteborg: Department of Computational Linguistics.

Ejerhed, E., Källgren, G., Wennstedt, O. and Åström, M. 1992. "The Linguistic Annotation System of the Stockholm-Umeå Corpus Project." Report 33. University of Umeå: Department of Linguistics.

Everitt, B. S. 1977. *The Analysis of Contingency Tables*. London: Chapman and Hall.

Gale, W. A. and Church, K. W. 1990. "Poor Estimates of Context Are Worse Than None." In *Proceedings of the Speech and Natural Language Workshop*, 283–287. Morgan Kaufmann.

Gale, W. A. and Sampson, G. 1995. "Good-Turing Frequency Estimation Without Tears." *Journal of Quantitative Linguistics* 2, 217–237.

Garside, R. 1987. "The CLAWS Word Tagging System." In Garside, R., Leech, G. and McEnery, A. (eds) *The Computational Analysis of English*. London and New York: Longman.

Garside, R. 1995. "Grammatical Tagging of the Spoken Part of the British National Corpus: A Progress Report." In Leech, G., Myers, G. and Thomas, J. (eds) *Spoken English on Computer: Transcription, Mark-up and Application*, 161–167. Longman.

Good, I. J. 1953. "The Population Frequencies of Species and the Estimation of Population Parameters." *Biometrika* 43, 45–63.

Jelinek, F. 1997. *Statistical Methods for Speech Recognition*. MIT Press.

- Jelinek, F. and Mercer, R. 1985. "Probability Distribution Estimation from Sparse Data." *IBM Technical Disclosure Bulletin* 28, 2591–2594.
- Katz, S. 1987. "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer." *IEEE Transactions on Acoustics, Speech and Signal Processing* 35, 400–401.
- Källgren, G. 1996. "Linguistic Indeterminacy as a Source of Errors in Tagging." In *Proceedings of the 16th International Conference of Computational Linguistics (COLING-96)*. Copenhagen: Center for Language Technology.
- Lidstone, G. J. 1920. "Note on the General Case of the Bayes-Laplace Formula for Inductive or *A Posteriori* Probabilities." *Transactions of the Faculty of Actuaries* 8, 182–192.
- Lindgren, B. W. 1993. *Statistical Theory*. Chapman-Hall.
- Merialdo, B. 1994. "Tagging English Text with a Probabilistic Model." *Computational Linguistics* 20, 155-171.
- Ney, H., Martin, S. and Wessel, F. 1997. "Statistical Language Modeling Using Leaving-One-Out." In Young, S. and Bloothoof, G. (eds) *Corpus-Based Methods in Language and Speech Processing*. Kluwer Academic Publishers.
- Nivre, J. 1999a. "Transcription Standard. Version 6." Göteborg University: Department of Linguistics. (Available at: [http://www.ling.gu.se/SLSA/Postscripts/Transcription\\_standard\\_v6.ps](http://www.ling.gu.se/SLSA/Postscripts/Transcription_standard_v6.ps)).
- Nivre, J. 1999b. "Modifierad standardortografi (MSO6)." Göteborg University: Department of Linguistics. (Available at: <http://www.ling.gu.se/SLSA/Postscripts/MSO6.ps>).
- Nivre, J. Forthcoming. "Sparse Data and Smoothing in Statistical Part-of-Speech Tagging." To appear in *Journal of Quantitative Linguistics*.
- Nivre, J., Allwood, J., Holm, J., Lopez-Kästen, D., Tullgren, K., Ahlsén, E., Grönqvist, L. and Sofkova, S. 1998. "Towards Multimodal Spoken Language Corpora: TransTool and SyncTool." In *Proceedings of the Workshop on Partially Automated Techniques for Transcribing Naturally Occurring Speech*, COLING-ACL '98, Montreal, Canada.
- Nivre, J., Grönqvist, L., Gustafsson, M., Lager, T. & Sofkova, S. 1996. "Tagging Spoken Language Using Written Language Statistics." In *Proceedings of the 16th International Conference of Computational Linguistics (COLING-96)*. Copenhagen: Center for Language Technology.
- Samuelsson, C. 1994. "Morphological Tagging Based Entirely on Bayesian Inference." In Eklund, R. (ed) *NODALIDA '93. Proceedings of the '9:e Nordiska Datalingvistikdagarna', Stockholm, 3–5 June 1993*. Stockholm University: Department of Linguistics.
- Svartvik, J. and Quirk, R. (eds) 1980. *A Corpus of English Conversation*. Lund Studies in English 56. Lund: Liber/Gleerups.
- Viterbi, A. J. 1967. "Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm." *IEEE Transactions on Information Theory* 13, 260–269.

## A The Large Tagset (\* = open part-of-speech)

<b>Tag</b>	<b>Part-of-speech</b>
ab*	adverb
dt	determiner
fb	feedback word
ha	<i>wh</i> -adverb
hd	<i>wh</i> -determiner
hp	<i>wh</i> -pronoun
hs	<i>wh</i> -possessive
ie	infinitive marker
in*	interjection
jj*	adjective
kn	conjunction (coordinating)
nn*	noun
ocm	own communication management
pc*	participle
pl	(verb) particle
pm*	proper name
pn	pronoun
pp	preposition
ps	possessive
rg*	cardinal numeral
ro	ordinal numeral
sn	subjunction
vb*	verb

## B The Small Tagset (\* = open part-of-speech)

<b>Tag</b>	<b>Part-of-speech</b>
adj*	adjective
adv*	adverb
fb	feedback word
int*	interjection
conj	conjunction
noun*	noun
num*	numeral
ocm	own communication management
pron	pronoun
prep	preposition
verb*	verb