Hybrid Techniques for Training HMM Part-of-Speech Taggers

Ted Briscoe
Greg Grefenstette
Lluís Padró
Iskander Serail

Report LSI-96-11-R

HYBRID TECHNIQUES FOR TRAINING HMM PART-OF-SPEECH TAGGERS

Ted Briscoe and Greg Grefenstette
Rank Xerox Research Centre
38240 Meylan, France
briscoe / grefen @xerox.fr

Lluís Padró
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Pau Gargallo 5, 08028 Barcelona, Spain
padro@lsi.upc.es

Iskander Serail
Computer Centrum Letteren
University of Amsterdam
Spuistraat 134, NL-1012 VB, The Netherlands
iskandar@alf.let.uva.nl

8 June 1994

Summary

We describe and experimentally evaluate a hybrid technique for training part-of-speech taggers which utilises training from small quantities of unambiguously-tagged material combined with maximum likelihood re-estimation over the target untagged corpus. This approach, unlike previous ones employing re-estimation, does not involve skilled manipulation of the initial parameters of the model or the use of sophisticated models of suffix-tag probabilities derived from unambiguously-tagged material. We conclude that this technique can yield usefully accurate taggers for several languages, but that the conditions required for success are difficult to state precisely.

Subject Areas: ANLP-94 Submission, N-gram/(H)MM Part-of-speech tagging, Maximum Likelihood Re-estimation

Word Count: 3800 approx

HYBRID TECHNIQUES FOR TRAINING HMM PART-OF-SPEECH TAGGERS

Summary

We describe and experimentally evaluate a hybrid technique for training part-of-speech taggers which utilises training from small quantities of unambiguously-tagged material combined with maximum likelihood re-estimation over the target untagged corpus. This approach, unlike previous ones employing re-estimation, does not involve skilled manipulation of the initial parameters of the model or the use of sophisticated models of suffix-tag probabilities derived from unambiguously-tagged material. We conclude that this technique can yield usefully accurate taggers for several languages, but that the conditions required for success are difficult to state precisely.

Subject Areas: ANLP-94 Submission, N-gram/(H)MM Part-of-speech tagging, Maximum Likelihood Re-estimation

Word Count: 3800 approx

1. BACKGROUND

Part-of-speech tagging is a practical, robust technique for assigning the unique, appropriate part-of-speech (or lexical syntactic category or simply 'tag') drawn from a predefined tagset to a word in context, by combining lexical information concerning the probability of a word taking a particular tag with contextual information concerning the probability of contiguous sequences of two or three tags (bigrams or trigrams). Taggers with low error rates have been usefully deployed in information retrieval and robust parsing systems, sometimes in conjunction with thresholding to avoid residual errors (de Marcken, 1990). The basic technique was developed at Lancaster University to annotate the Lancaster-Oslo/Bergen corpus (Leech *et al*, 1983; Garside, 1987). Both lexical and transition probabilities are estimated from a training corpus of manually tagged material. A more efficient dynamic programming version of the technique was presented independently by Church (1988) and de Rose (1988). Both these authors report success rates above 95% (of correct assignment of tags to words in context) based on self-tests (training and testing) on the Brown Corpus. The relationship between these models and Markov modelling techniques has been pointed out by several authors. In particular, the linear-time dynamic programming techniques utilising bigrams or trigrams and training on unambiguous data are special cases of Markov models utilising the Viterbi algorithm (Viterbi, 1967; Jelinek *et al*, 1975).

Recently, the Baum-Welch algorithm (Baum, 1972), a self-organising technique for 'hidden' Markov models which can be applied to ambiguous training data (Jelinek, 1985) has been utilised for part-of-speech tagging (Kupiec, 1992; Cutting *et al*, 1992). In this latter model, lexical probabilities for all but the most frequent words are replaced by equivalence

classes of words assigned the same ambiguous set of tags, and bigram transition probabilities are augmented with manually-specified networks of 'tied' states to overcome observed common errors in the bigram model, effectively creating a partly second-order model. The model is trained by manually biasing initial probabilities of bigrams and of tags in ambiguous sets. The Baum-Welch/Forward-Backward algorithm allows polynomial-time re-estimation of these probabilities by iteratively modifying them to maximise the likelihood of the training corpus given the probabilistic model. The trained model is then applied as a linear-time tagger using the Viterbi algorithm. Training on one-half of the Brown Corpus and testing on the other gave a success rate of over 96%, suggesting that training from ambiguous untagged data using maximum likelihood (ML) re-estimation can, in principle, yield results as good as direct estimation of parameters from unambiguously-tagged training data.

This result is of potential practical import since large manually-corrected unambiguously-tagged corpora currently only exist for a few languages, and the construction of such corpora is labour intensive. Meteer *et al* (1991) estimate that around 70K words of such training material is required to produce an accurate trigram tagger if direct estimates of parameters are required (given the extreme bias in distribution of observed trigrams), whilst Cutting *et al* (1992) mention that they have achieved reasonable results using iterative re-estimation on 3000 untagged sentences. However, both these lower-limit estimates of training corpus size assume that some other accurate method is employed to estimate lexical probabilities or ambiguity class membership for unknown words. Meteer *et al* (1991) demonstrate that a trigram model which estimates lexical probabilities for unknown words from the probability of a tag occurring with a suffix in unambiguous training data can achieve an 82% success

rate on assigning the correct tag in context to unknown words. Kupiec (1992) utilises a very similar approach to induce biases in the relative probabilties of tags in the ambiguity class for unknown words. Unfortunately, the estimation of suffix-tag probabilities from unambiguous data re-introduces the requirement for unspecified quantities of tagged training data and undermines the claim that new languages can be accurately tagged automatically.

ML re-estimation techniques converge to a local optimum maximising the overall probability of the training corpus, (expressed as the product of the sums of the possible paths used to generate each sentence of the corpus), but there is no *a priori* reason to suppose that the derived model will assign highest probability to the linguistically correct path of tags for each sentence. It is, therefore, highly likely that initial biases in ambiguity classes and transition probabilities are essential to guide the model to the linguistically desirable model. Merialdo (1991) reports several tagging experiments utilising ML re-estimation on ambiguous training material which demonstrate that utilising a trigram model with lexical probabilities and starting with random initial probabilities yields less accurate results than training the same model on as little as 100 sentences of unambiguous material. He also demonstrates that further training of a model with accurate directly estimated parameters degrades performance. Elworthy (1993, submitted) refines these results confirming that the utility of ML re-estimation is a function of the 'similarity' between the model and corpus and the 'conservativity' of re-estimation attempted. These results suggest that it will be difficult if not impossible to predict the utility of re-estimating from an initial model without access to an unambiguous corpus to assess the 'goodness-of-fit' of the model and evaluate the effect of re-estimation.

2. OUR GOALS

The results summarised in §1 suggest that the combined problems of assessing the effect of re-estimation and dealing accurately with unknown words create the need for unspecified quantities of unambiguously-tagged material for construction of an accurate tagger, even when ML re-estimation is deployed. In this paper, we explore the utility of ML re-estimation with a bigram model incorporating lexical probabilities as a technique for improving an approximate model derived from very limited unambiguous data. Our overall goal is the development of a viable methodology for accurately tagging languages for which no prior unambiguous training corpora are available. The specific questions we address are:

1. Can useful initial model parameters be automatically acquired from a small unambiguously-tagged training corpus of approximately 10K words?

2. Can these initial parameters be refined to obtain an accurate model using ML re-estimation over a larger ambiguous untagged corpus?

3. Can the unknown word problem be reduced to re-estimation of (lexical) probabilities from the large ambiguous corpus?

4. Can a bigram model incorporating lexical probabilities produce accurate taggers for languages which exhibit different trade-offs between inflectional morphological complexity and word order restrictions?

5. Can we know when we have obtained an accurate model in the absence of a large unambiguously-tagged corpus?

3. THE TAGGING SYSTEM AND EXPERIMENTAL DESIGN

We utilise the Acquilex Tagger (Elworthy, 1993) in the experiments described below. This system implements both the Viterbi and Forward-Backward/Baum-Welch algorithms for a bigram model incorporating lexical probabilities and allows both direct estimation of probabilities from unambiguous training data and iterative re-estimation using ambiguous data. The system does not incorporate a morphological analyser so dictionaries of word forms with associated tags were built from tagged corpora, except in the case of French where no prior tagged data was available and a dictionary was created by assigning all possible tags to each distinct word form utilising a finite-state lexical transducer for French (Karttunen *et al*, 1992; Chanod, submitted). The tagger assigns all open-class tags with equiprobability to unknown words whilst performing tag disambiguation.

This tagger was trained with 6 tagsets ranging in size from 10 to 317 tags and was applied to English, French, Spanish and Dutch data. The English data utilised was the Susanne Corpus (Sampson, in press), a 150K word tagged and parsed subset of the Brown Corpus. We extracted from this corpus an unambiguously-tagged corpus utilising 230 tags very similar to CLAWS-II (Garside *et al*, 1987) without idiom tags. The French corpus utilised was a 15M corpus of untagged newspaper text from *Libération*. This corpus was ambiguously tagged with a set of 268 tags using a morphological analyser. The Dutch Eindhoven Corpus contains 368K words of newspaper, magazine and other non-specialist text which has been manually tagged with a tagset of 317 tags, including 149 subcategories of open-class parts-of-speech (den Boogaart, 1975). A small (10 tag) version of this tagset was

obtained by suppressing most of the distinctions in the big tagset. The Spanish Biblograf-Vox Corpus consisted of 6M words of text from various sources of which a 17K sample has been tagged and hand corrected with a tagset of 28 parts-of-speech, incorporating tense distinctions amongst verbs (Moreno-Torres, 1994). Elworthy (1993) demonstrates that there is no straightforward correlation between tagset size, membership or the average word/tag ambiguity rate and tagging accuracy, making comparison and evaluation of some of the results reported below more difficult.

For each of the experiments reported below we provide the following measures: overall percentage of correctly tagged words, percentages of correctly tagged known and unknown words (where appropriate), percentages of all correctly tagged ambiguous words and of known ambiguous words, and the perplexity for the corpus after each iteration of re-estimation (see e.g. Sharman, 1990). These measures allow more accurate assessment of the performance of the tagger than the overall accuracy.

4. THE ENGLISH EXPERIMENTS

We used the Susanne Corpus to explore the possibility of building an accurate model from limited unambiguously-tagged data and to refine the methodology for achieving this goal.

4.1. Training with a Large Corpus

In order to assess the best performance of the tagger, we trained on the entire Susanne Corpus using both the Viterbi and Forward-Backward (FB) algorithms. We then applied re-estimation to the resulting system. The results of self-tests on the entire corpus demonstrate

All words (Viterbi)	97.77%
All ambiguous words (46.04%)	95.17%
All words (FB)	97.80%
All ambiguous words (46.04%)	95.23%

Figure 1: Self-Tests Without Re-estimation

All words	97.25%	96.62%	95.85%
All ambig. words (46.04%)	94.03%	92.67%	90.98%
Perplexity	-8.07015	-8.06027	-8.04931

Figure 2: Self-Tests With Re-estimation: 2,5 and 10 Iterations

that the tagger is capable of better than previously reported self-test performance (with this tagset on this corpus overall accuracy 97.8%), see Figure 1. However, when ML re-estimation is applied to this accurate model it degrades performance, replicating the results of Elworthy (1993, submitted) and Merialdo (1991) with a different tagset and corpus, and underlining the need for care in the use of ML re-estimation techniques; Figure 2 clearly indicates that as ML re-estimation reduces perplexity, the accuracy of tagging decreases.

4.2. Training with a Small Corpus

A small amount of unambiguously-tagged training data might provide enough information to directly estimate transition probabilities and very frequent (mostly closed-class) lexical probabilities such that further re-estimation from the resulting model would produce a useful tagger. In effect, this approach would substitute manual tag disambiguation for manual work biasing initial probabilities in a model based completely on re-estimation (Cutting *et al*, 1992). This could be advantageous if small quantities of tagged data were already available and/or the skill required to manually disambiguate tags were more common than

All words (FB)	77.84%
Known words	92.04%
Unknown words (23.65%)	31.99%
All ambiguous words (53.38%)	61.79%
Known ambiguous words	85.50%

Figure 3: Small Training Corpus Without Re-estimation

All words (FB)	79.93%	80.35%	79.58%
Known words	92.47%	92.83%	91.95%
Unknown words (23.65%)	39.48%	40.08%	39.66%
All ambiguous words (53.38%)	65.71%	66.50%	65.06%
Known ambiguous words	86.58%	87.52%	85.27%
Perplexity	-10.4604	-9.17922	-8.14802

Figure 4: Small Training Corpus With Re-estimation; 2,3 and 5 Iterations

that required to set ambiguity class biases for a specific type of tagger.

In Figure 3 we report the results of training the tagger on 10K words of tagged material using the FB algorithm and using the resulting model to tag the remainder of the corpus. The overall performance is bad, and that on unknown words worse, in line with previous results assigning unknown words all open-class tags equiprobably (Meteer *et al*, 1991). This system was also used as the starting point for re-estimation over the remainder of the corpus. The results are given in Figure 4. ML re-estimation improves performance slightly, but peaks in the third iteration and afterwards tagging accuracy decreases, though perplexity continues to be reduced. Thus, ML re-estimation can improve the performance of an approximate model constructed from small quantities of unambiguously-tagged material. However, the best performance of the re-estimated model (80.35%) is not useful. It is clear from the results that poor performance stems to a large extent from the lack of a sophisticated model for

All words	93.51%
All ambiguous words (46.04%)	85.89%

Figure 5: Ideal Dictionary Without Re-estimation

unknown words. Furthermore, re-estimation does not directly improve the performance on unknown words since only lexical and transition probabilities acquired from the 10K training material are re-estimated.

4.3. Adding Ideal Lexical Information

In order to assess the potential contribution of accurate lexical probabilities to tagging accuracy, we merged the transition table derived from the model directly-estimated from 10K words with the dictionary derived from the original self-test. The results of tagging the remainder of the corpus are given in Figure 5, providing a ceiling against which to assess our attempts to infer useful approximate models for lexical probabilities. Accurate lexical probabilities reduce the error rate on the best model in §4.2 by a factor of three.

4.4. Adding Approximate Lexical Information

The existing techniques for improving accuracy on unknown words rely on unambiguously-tagged material from which to acquire statistics concerning suffix-tag probabilities (Meteer *et al*, 1991; Kupiec, 1992). Placing known words into equivalence classes of ambiguous sets of tags does not solve this problem because unknown words will remain associated with the largest such set of open class tags. Given the availability of wide-coverage and efficient text tokenisers (Grefenstette and Tapanainen, in press) and morphological analysers (e.g. Ritchie *et al*, 1992; Karttunen *et al*, 1992) it is possible in principle to assign all the morphologi-

All words (FB)	92.92%	93.81%	92.87%
All ambiguous words (53.38%)	84.59%	86.54%	84.50%
Perplexity	-36.4782	-12.2636	-11.6188

Figure 6: Equiprobable Dictionary; 1,3 and 40 Iterations

cally appropriate tags to a very high percentage of words in running text. Such systems do not provide lexical probabilities, but the unknown word problem can be circumvented by approximating all lexical probabilities and using ML re-estimation over the target corpus to refine them.

We simulated the output of an appropriate morphological analyser for English by taking the existing exhaustive dictionary built in the self-test and modifying the lexical probabilities in the first case to be equiprobable and in the second to include a uniform weighting based on the frequency of the tag in the 10K word corpus. We then combined these two dictionaries with the dictionary derived from the 10K word corpus, so that lexical probabilities for high frequency words were more accurate, and we also used the transition probabilities derived from this small corpus. The two resulting models were re-estimated over the remainder of the untagged corpus. The results for each model are given in Figures 6 and 7. In both cases, a 100 iterations were used: perplexity did not decrease further beyond about 50 iterations at -10.0299 for the frequency weighted model and at -11.6133 for the equiprobable model. In the frequency weighted model, tagging accuracy stabilised after the 25th iteration improving (marginally) to 93.05% overall accuracy at the 100th iteration. In the equiprobable model, tagging accuracy peaked on the third iteration at 93.81% overall accuracy and stabilised around 92.87% after 40 iterations.

All words (FB)	88.91%	93.01%	93.05%
All ambiguous words (53.38%)	75.75%	84.70%	84.79%
Perplexity	-33.0941	-10.0352	-10.0243

Figure 7: Frequency-Weighted Dictionary; 1,40 and 100 Iterations

Thus, the combination of a wide-coverage and accurate morphological analyser and a 10K word unambiguously-tagged corpus combined with ML re-estimation over the target corpus can yield a bigram model with a useful performance, though the error rate for the derived models is three times greater that that for the original self-test and around twice as bad as the results obtained by Cutting *et al* (1992) when re-estimation was supplemented with a directly estimated model for unknown words. Meaningful comparison with the latter result is difficult as Cutting *et al* did not report a result for training without re-estimation and it is unclear how their use of ambiguity classes and a smaller tagset will influence overall performance. Nevertheless, the performance of our derived models on ambiguous words improves on the 82% performance reported by Meteer *et al* (1991) using suffix-tag probabilities for unkown words, suggesting that re-estimation over the target corpus is an effective alternative to unknown word strategies requiring access to an untagged corpus.

5. THE FRENCH EXPERIMENT

In order to further assess the approach developed above, we assigned all possible tags to each word of a 15M word sample of French newspaper text using a finite-state tokeniser (Grefenstette and Tapanainen, in press) and lexical transducer (Chanod, submitted). The tagset employed contained 268 tags, however it did not incorporate semantic subclasses of nouns, and the additional size was created by additional verb tense and gender distinctions.

All words (Self-Test)	98.94%
All ambiguous words (23.70%)	95.53%
All words (Unseen Test)	98.44%
All ambiguous words (22.07%)	94.77%

Figure 8: French Self and Unseen Test after Re-estimation

For this reason, the average ambiguity of words is lower than for the English corpus (around 1.2 tags/word as compared to 1.4 tags/word).

We hand-disambiguated 10.5K words of this corpus which was split into a 8.5K training corpus and 2K test corpus. A tagger was constructed using direct estimation of transition and lexical probabilities from frequencies in the hand-tagged training corpus. Probabilities of unseen word-tag combinations for all words in the 15M corpus were estimated initially by assuming all such combinations had been seen once, as in the equiprobable model of §4.4. This model was then re-estimated using 10 iterations and used to retag the unambiguous training corpus and the test data. The results are given in Figure 8. These results are promising in that they suggest that our approach can yield an accurate model and that French is amenable to accurate tagging utilising bigrams and lexical probabilities. However, the results should be treated with some caution, as the lack of a substantial tagged corpus makes serious tests of the accuracy of the resulting tagger difficult. Assuming the performance is as good as indicated it is possible that the properties of the tagset, the use of a much larger corpus for re-estimation, and/or the inherent complexity of the corpus have contributed to the improvement in performance over the comparable English experiments.

All words (Self-Test)	94.89%
All ambiguous words (58.53%)	91.31%
All words (Unseen Test)	84.61%
Known words	91.76%
All ambiguous words (72.78%)	80.15%
Known ambiguous words	89.63%

Figure 9: Dutch Tests with Large Tagset

6. THE DUTCH EXPERIMENT

Similar experiments were undertaken for Dutch, first using the tagged corpus to train the tagger with an extended tagset encoding morphological and limited functional information (317 tags). The results of a self-test on the entire corpus and an unseen test training on two thirds of the corpus and testing on another third with the FB algorithm are given in Figure 9. They demonstrate that a bigram model with lexical probabilities is capable of yielding a useful tagger for Dutch. However, performance on known ambiguous words declines more than might be expected in the unseen test, suggesting that the tagset may not be optimal for HMM techniques. (The tagset was developed for entirely manual tagging – den Boograat, 1975.)

In order to see whether a more conventional tagset would yield a better result a reduced set of 10 tags was derived encoding only basic part-of-speech. The results for the equivalent experiments are shown in Figure 10. It is probable that with an optimal tagset (removing morphological information is known to degrade performance for English, Elworthy, 1993) a better model could be obtained.

To test the efficacy of our approach for constructing models from small training corpora the tagger was trained on 10K of tagged text for both tagsets and the resulting dictionary

All words (Self-Test)	97.68%
All ambiguous words (35.18%)	93.40%
All words (Unseen Test)	90.54%
Known words	96.54%
All ambiguous words (51.91%)	82.42%
Known ambiguous words	93.97%

Figure 10: Dutch Tests with Small Tagset

All words (Small Tagset)	96.04%	95.19%
All ambiguous words (35.19%)	88.76%	86.33%
Perplexity	-8.37562	-9.23945
All words (Big Tagset)	83.57%	88.53%
All ambiguous words (56.10%)	73.18%	81.27%
Perplexity	-61.7728	-13.8405

Figure 11: Dutch Unseen Test with Re-estimation; Iterations 1 and 10

was merged with an equiprobable one derived from the remainder of the tagged corpus, as in §4.4. The best results after re-estimating over the remaining untagged corpus for both tagsets are shown in Figure 11. These suggest that the model obtained with the small tagset was too good for re-estimation to achieve any improvement, which is perhaps not surprising given the small number of transition parameters. The result for the large tagset indicates that re-estimation yields an improvement, but does not lead to a useful model. Further experimentation with an optimal tagset will be required to determine whether our approach a usefully accurate model.

7. THE SPANISH EXPERIMENT

Similar experiments were conducted for Spanish. Preliminary results show that the tagger produces an ambiguous word rate of 95.52% on a self-test on the tagged 17K word corpus and

85.09% for an unseen test using a merged dictionary with equiprobable lexical probabilities for unkown words after training on 10K tagged words. The performance of this model using re-estimation over the much larger untagged corpus will be reported in the final version of the paper.

8. CONCLUSIONS AND FURTHER RESEARCH

The experimental results we have obtained suggest that under certain conditions it is possible to construct accurate part-of-speech taggers utilising bigrams in conjunction with lexical probabilities by using a hybrid approach, training on very small unambiguously-tagged corpora and using ML re-estimation to improve the accuracy of the directly-estimated model. Furthermore, the results suggest that accurate bigram tagging of several languages with less fixed word order than English is possible. Therefore, we can answer the first four of the five questions of §2 positively. However, further work is required to identify the conditions in which ML re-estimation will improve accuracy in the absence of a tagged corpus and to identify the upper limit to the accuracy of the models which can be acquired with our training methodology. The lack of any straightforward relationship between perplexity and tagging accuracy suggests that this may be difficult. Furthermore, for French, Dutch and Spanish more definitive experiments must await the availability of tagged corpora with optimal tagsets for HMM techniques.

PRIVATE CONCLUSIONS (NOT IN SUBMISSION!)

The following further work should be done regardless of acceptance or not of the paper, so that this paper can be improved and a longer version can be submitted to *Computer*

Speech and Language:

1. Redo the French experiment so that perplexity and accuracy at various iterations is recorded against the 2K test data

2. Work out unigram results for the various trained models – good indication of the goodness of lex. probs. acquired

3. Get a unigram result for Xerox tagger on Brown corpus

4. Produce a better tagset for Dutch by more careful collapsing of the large one, so that morph subclasses are preserved + redo exps.

5. Do a proper test with a morph analyser for Eng – where the tags returned are appropriate for the tagger?

6. Merge this paper with David Elworthy's ANLP94 submission to produce a joint one

7. Include some discussion of tagset effects from Elworthy 1993

ACKNOWLEDGEMENTS

This research was partially supported by EC Esprit BR Project 7315 'The Acquisition of Lexical Knowledge' (Acquilex-II).

REFERENCES

Baum, L. (1972). "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes." *Inequalities,* III: 1–8.

den Boogaart, P. (1975). *Woordfrequenties.* Utrecht: Oosthoek, Scheltema en Holkema.

Chanod, J.P. (submitted). "Finite-state composition of French verb morphology." In *Proceedings, 4th Conf. Applied NLP.*

Church K (1988). "A stochastic parts program and noun phrase parser for unrestricted text." In *Proceedings, 2nd Applied ACL.* 136–143.

Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. (1992). "A practical part-of-speech tagger." In *Proceedings, 3rd Applied ACL.* 133–140.

Elworthy, D. (1993). *Part-of-Speech Tagging and Phrasal Tagging.* Acquilex-II Working Paper 10, Cambridge University Computer Laboratory†.

Elworthy, D. (submitted). "Does Baum-Welch re-estimation help taggers?." In *Proceedings, 4th Conf. Applied NLP.*

Garside, R. (1987). "The CLAWS word-tagging system." In *The Computational Analysis of English: A Corpus-based Approach*, edited by R. Garside, G. Leech and G. Sampson, 30–41. London, UK: Longman.

Garside, R., Leech, G., and Sampson, G. (1987). *The Computational Analysis of English: A Corpus-based Approach.* London, UK: Longman.

Grefenstette, G. and Tapanainen, P (in press). "What is a word?; what is a sentence?: problems of tokenisation." In *Proceedings, COMPLEX-94.*

Jelinek, F. (1985). "Markov source modelling of text generation." In *Impact of Processing Techniques on Communication*, edited by J. Skwirzinski, Dordrecht: Nijhoff.

Jelinek, F., Bahl, L and Mercer, R. (1975). "Design of a linguistic statistical decoder for the recognition of continuous speech." *IEEE Trans. on Information Theory,* IT-21: 250–256.

Karttunen, L., Kaplan, R., & Zaenen, A. (1992). "Two-level morphology with composition." In *Proceedings, COLING-92.* 141–148.

Kupiec, J. (1992). "Robust part-of-speech tagging using a hidden Markov model." *Computer Speech and Language,* 6: 1–21.

de Marcken, C. (1990). "Parsing the LOB corpus." In *Proceedings, 28th Annual Meeting of the Association for Computational Linguistics.* 243–251.

Merialdo, B. (1991). "Tagging text with a probabilistic model." In *Proceedings, ICASSP-91.* 809–812.

Meteer, M., Schwartz, R., and Weischedel, R. (1991). "POST: Using probabilities in language processing." In *Proceedings, 12th International Joint Conference on Artificial Intelligence.* 960–965.

Moreno-Torres, I. (1994). *A morphological disambiguation tool: application to Spanish.* Acquilex-II Working Paper 24, Universitat Politécnico di Catalunya†.

Ritchie, G., Russell, G., Black, A. & Pulman, S. (1992). *Computational Morphology.* MIT Press.

de Rose, S. (1988). "Grammatical category disambiguation by statistical optimization." *Computational Linguistics,* 14.1: 31–39.

Sharman, R (1990). *Hidden Markov Model Methods for Word Tagging.* UKSC-TR-214, IBM UK Scientific Centre, Winchester, England.

Viterbi, A. (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm." *IEEE Transactions on Information Theory,* IT-13: 260–269.

†Acquilex-II working papers can be obtained by sending a request to `cide@cup.cam.ac.uk`

Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

Research Reports - 1996

LSI-96-1-R "(Pure) Logic out of Probability", Ton Sales.

LSI-96-2-R "Automatic Generation of Multiresolution Boundary Representations", C. Andújar, D. Ayala, P. Brunet, R. Joan-Arinyo, and J. Solé.

LSI-96-3-R "A Frame-Dependent Oracle for Linear Hierarchical Radiosity: A Step towards Frame-to-Frame Coherent Radiosity", Ignacio Martin, Dani Tost, and Xavier Pueyo.

LSI-96-4-R "Skip-Trees, an Alternative Data Structure to Skip-Lists in a Concurrent Approach", Xavier Messeguer.

LSI-96-5-R "Change of Belief in SKL Model Frames (Automatization Based on Analytic Tableaux)", Matías Alvarado and Gustavo Núñez.

LSI-96-6-R "Compressibility of Infinite Binary Sequences", José L. Balcázar, Ricard Gavaldà, and Montserrat Hermo.

LSI-96-7-R "A Proposal for Word Sense Disambiguation using Conceptual Distance", Eneko Agirre and German Rigau.

LSI-96-8-R "Word Sense Disambiguation Using Conceptual Density", Eneko Agirre and German Rigau.

LSI-96-9-R "Towards Learning a Constraint Grammar from Annotated Corpora Using Decision Trees", Lluis Màrquez and Horacio Rodríguez.

LSI-96-10-R "POS Tagging Using Relaxation Labelling", Lluís Padró..

LSI-96-11-R "Hybrid Techniques for Training HMM Part-of-Speech Taggers", Ted Briscoe, Greg Grefenstette, Lluís Padró, and Iskander Serail.

Hardcopies of reports can be ordered from:

Nuria Sánchez
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Pau Gargallo, 5
08028 Barcelona, Spain
secrelsi@lsi.upc.es

See also the Department WWW pages. http://www-lsi.upc.es/www/