Some Experiments with the CRATER Corpus

Thorsten Brants
Universität des Saarlandes, Computational Linguistics
P.O.Box 151150, D-66041 Saarbrücken, Germany
thorsten@coli.uni-sb.de

November 6, 1995

Report on Corpus Applications for the CRATER Project

Abstract

This paper reports statistical information about the CRATER corpus and experiments performed with it. The corpus is compiled, part-of-speech annotated and manually edited by the Corpus Resources and Terminology Extraction project. The experiments concern statistical part-of-speech tagging and statistically motivated tagset modification.

1 Introduction

One application for statistical language models is part-of-speech tagging. Here, the task is the unique annotation of a word with a syntactic category, called *part-of-speech* or *tag*. There are two different main methods for deriving language models. The first one uses a previously (manually) annotated corpus and estimates the model parameters from frequencies of tags, frequencies of combinations of tags, and frequencies of combinations of tags and words in the training corpus. The second method uses a plain corpus, a lexicon stating the possible parts-of-speech for each word, a prior bias for the model parameters and the Baum-Welch algorithm [Baum *et al.*, 1970].

[Elworthy, 1994] shows that the success of the Baum-Welch algorithm depends heavily on the bias which has to be chosen in advance, and generally the first method yields better results. But the construction of a manually annotated corpus is very time consuming, so for practical reasons often the second method is chosen.

The situation has now changed for English, French, and Spanish. The CRATER project (Corpus Resources and Terminology Extraction) compiled, annotated with parts-of-speech and manually corrected a large parallel corpus for these three languages (470,000 to 760,000 words for each language). It is composed of telecommunications texts provided by the International Telecommunications Union in Geneva. The purpose of the corpus is to prepare a high quality trilingual parallel aligned text for the NLP research community.

The following sections report data about the CRATER corpus and experiments performed with it. These experiments concern statistical part-of-speech tagging and statistically motivated tagset modification.

2 Experiments with the English Corpus

2.1 Basic Information about the Corpus

The English part of the CRATER corpus is the version as of April 4th, 1995. It consists of 69 files, named EN000 to EN070, missing EN006 and EN013. Each file contains between 2,223 (EN033) and 46,421 (EN059) tokens (words, numerals, punctuation marks, special characters). All files together contain 764,677 tokens.

The corpus is annotated using the CLAWS tagset C7 which consists of 152 tags. Of these, 141 tags were actually used. Each tag can be indexed to mark multi-word lexemes (e.g., *instead of* is tagged CC21 and CC22). When counting indexed tags as separate tags, 217 tags where used in total.

There are 20,760 different tokens (types) in the corpus, not distinguishing upper and lower case letters. Table 1 shows the 10 most frequent tokens in the corpus. The by far most frequent token is the word *the* with 51,332 occurrences, which is 6.8% of the corpus (every 15th token is *the*).

Table 2 shows the 5 most frequent tags. The most frequent tag is NN1 (singular common noun), which occurs 170,955 times or 22% of the corpus.

Table 3 shows the number of tags per token in the corpus. This indicates the amount of work for a part-of-speech tagger. There is nothing do for tokens with only one possible tag, which happens for 26.79% of the corpus. For the rest (73.71%), disambiguation is necessary. The average number of tags per token is 3.47. This is a high average compared with the Susanne Corpus [Sampson, 1995] which uses 424 tags (incl. indexed tags for multi-word lexemes) and has an average number of 2.61 tags per token, or compared with the Spanish part of the CRATER corpus, which uses 315 tags and has an average number of 1.75 tags per token. Table 4 shows the 5 most ambiguous words in the English part of the CRATER corpus.

2.2 Statistical Trigram Tagging

This section reports statistical trigram tagging experiments performed with the English part of the CRATER corpus. We used a standard statistical approach [Rabiner, 1989] and apply the Viterbi algorithm [Viterbi, 1967].

Sparse data is handled by estimated likelihood estimation (see e.g. [Gale and Church, 1990]): to avoid zero probabilities for transitions between tags we add 0.5 to each frequency before calculating the maximum likelihood estimation of transition probabilities.

Unknown words are handled by analyzing the suffixes of length 3 of the unknown words. If there are other words in the lexicon which have the same suffix, the distribution of tags for these words is used for the unknown word. If the suffix is not found in the lexicon, we use the distribution of words which occured exactly once in the training part.

Numbers are recognized separately. They are identified as a sequence of digits with optional '.' or ','. We assign the distribution of tags for all numbers found in the training part to the number in the test part.

For each tagging task, we used 68 files of the corpus for training and the remaining file for testing. The overall results are shown in figure 1. In average, 96.20% of the test part is tagged correctly. This percentage is en par with results reported for other tagging tasks. But since we used a large tagset (217 tags), generally the reported tagsets are much smaller (< 100), this result can be regarded as better than other statistical tagging results. One reason for the good result is the size of the corpus ($> 750,000$ words). There is no other investigation using a manually tagged corpus of this size. Other reasons may be the structure of the corpus or the tagset and should be subject to further investigation.

Figure 2 shows the percentage of unknown words in each of the files of the corpus. Unknown words are those words that occur only in the test file and not in one of the other files. Figures 3 and 4 show the tagging results separately for known and unknown words.

2.3 Tagset Reduction

Generally, the categories for the tagging task are linguistically motivated and do not reflect probability distributions or co-occurrence probabilities of words belonging to the categories. Often, information not needed for probability estimation is contained in the tagset. E.g., distributions of comparative and superlative adjectives tend to be very similar, and combining these tags could increase the accuracy of probability estimation since there is more data for both tags together than for the single tags. But one is interested in the information encoded in each single tag, so one has to ensure that all information can be extracted from the combined (reduced) tagset. For details on the method see [Brants, 1995b].

The CLAWS tagset used for the experiments in the previous section can be reduced drastically without loosing information. The original tagset has 217 tags, the reduced one only 53, but the performance stayed almost constant. The resulting model has about $53^3 \approx 1.5 \cdot 10^5$ transition probabilities opposed to $217^3 \approx 10^7$ for the original model. Figure 5 shows the tagging accuracy after each step of tag clustering. Table 5 show the tags involved in the first 10 merges.

3 Experiments with the Spanish Corpus

3.1 Basic Information about the Corpus

The Spanish part of the CRATER corpus is the version as of September 26th, 1995. It consists of 56 files, named SP000 to SP054 and SP064. Each file contains between 1,408 (SP054) and 18,384 (SP002) tokens (words, numerals, punctuation marks, special characters). All files together contain 473,847 tokens.

The corpus is annotated using the tagset describe in [León, 1994] which consists of 492 tags. Of these, 315 were actually used.

Table 6 shows the 10 most frequent tokens in the corpus, not distinguishing upper and lower case letters. The by far most frequent token is the word *de* with 39,958 occurrences, which is 8.4% of the corpus (every 12th token is *de*).

Table 7 shows the 5 most frequent tags. The most frequent tag is PREP (preposition), which occurs 69,038 times or 14.6% of the corpus (every 7th token is a preposition).

Table 8 shows the number of tags per token in the corpus. This indicates the amount of work for a part-of-speech tagger. There is nothing do for tokens with only one tag, which happens for 55.91% of the corpus. For the rest (44.09%), disambiguation is necessary. The average number of tags per token is 1.75. Table 9 shows the 5 most ambiguous words in the corpus.

3.2 Statistical Trigram Tagging

This section reports statistical trigram tagging experiments performed with the Spanish part of the CRATER corpus. We used the same tagger as for the English experiments, sparse data and unknown words are handled identically.

For each tagging task, we used 55 files of the corpus for training and the remaining file for testing. The overall results are shown in figure 6. In average, 97.07 of the test part is tagged correctly. This percentage is very high compared with other tagging tasks. The tagset used is very large ($\sim$ 300 compared to $\sim$ 100 found in other investigations, or $\sim$ 200 for the English part of this corpus). On the one hand, this makes the tagging task more difficult, since the tagger has to find the correct tag from a larger set. But on the other hand, this makes the tagging task easier, since a lot of words have their own tags resulting in no ambiguity for these words. The specialized tagset causes the low average of 1.75 possible tags per token in running text.

Figure 7 shows the percentage of unknown words in each of the files of the corpus. Unknown words are those words that occur only in the test file and not in one of the other files. Figures 8 and 9 show the tagging results separately for known and unknown words.

4 Conclusion

The tagging experiments reported in this paper can serve as a baseline for tagging tasks. Handling sparse data and unknown words can be improved by using deleted interpolation [Brown *et al.*, 1992] instead of estimated likelihood estimation, and using decision trees for unknown word suffixes [Samuelsson, 1993] instead of suffixes of fixed length for unknown words.

Despite of big differences in the tagset tagging results for the English part are en par with results reported in investigations of other corpora. The high tagging accuracy is partly due to the large amount of training material. The results for the Spanish part are better than those for the English part. This should be mainly due to the specialized tagset.

The English tagset can be drastically reduced without loosing any information and without loosing accuracy.

Further work with the corpus will include the tagset reduction for the Spanish part. Additionally, it is intended to perform all reported experiments with the French part of the corpus.

All experiments are part of the work for a PhD, investigating the topology of Markov Models, i.e., the number of states for the model and for each state the outputs and transitions with non-zero probability. Generally, the topology reflects the tagset. This could be improved by making the states independent of the chosen tags.

The experiments reported in [Brants, 1995a] will be performed with the CRATER corpus.

5 Acknowledgment

I would like to thank Tony McEnery for supplying the CRATER corpus, and the whole team of the CRATER project for their work.

References

[Baum *et al.*, 1970] Leonard E. Baum, Ted Petrie, George Soules and Norman Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions in Markov chains. *The Annals of Methematical Statistics*, 41:164–171, 1970.

[Brants, 1995a] Thorsten Brants. Estimating HMM Topologies. In *Tbilisi Symposium on Language, Logic, and Computation*, Human Communication Research Centre, Edinburgh, HCRC/RP-72, 1995.

[Brants, 1995b] Thorsten Brants. Tagset Reduction Without Information Loss. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA., 1995.

[Brown *et al.*, 1992] P. F. Brown, V. J. Della Pietra, Peter V. deSouza, Jenifer C. Lai and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

[Elworthy, 1994] David Elworthy. Does Baum-Welch Re-estimation Help Taggers? In *Proceedings of ANLP-94*, 1994.

[Gale and Church, 1990] W. A. Gale and K. W. Church. Poor Estimates of Context are Worse than None. In *Proc. of the Speech and Natural Language Workshop*, pages 283–287, Hidden Valley, PA, 1990.

[León, 1994] Fernando Sánches León. *Spanish tagset for the CRATER project.* Technical Report Technical Report, Facultad de Filosofía y Letras, Madrid, 1994.

[Rabiner, 1989] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77(2), pages 257–285, 1989.

[Sampson, 1995] Geoffrey Sampson. *English for the Computer.* Oxford University Press, Oxford, 1995.

[Samuelsson, 1993] Christer Samuelsson. Morphological Tagging Based Entirely on Bayesian Inference. In *9th Nordic Conference on Computational Linguistics*, Stockholm University, Stockholm, Sweden, 1993.

[Viterbi, 1967] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. In *IEEE Transactions on Information Theory*, pages 260–269, 1967.

Table 1: 10 most frequent tokens in the English part of the CRATER corpus, not distinguishing upper and lower case letters.

	Token	Frequency			Token	Frequency	
1.	the	51,332	(6.7%)	6.	a	14,948	(2.0%)
2.	.	24,960	(3.3%)	7.	to	14,538	(1.9%)
3.	of	24,000	(3,1%)	8.	in	13,465	(1.8%)
4.	,	22,726	(3.0%)	9.	and	12,553	(1.6%)
5.	)	15,253	(2.0%)	10.	(	11,596	(1.5%)
				Σ		205,371	(26.9%)

Table 2: 5 most frequent tags in the English part of the CRATER corpus.

	Tag	Frequency		Description
1.	NN1	170,955	(22.4%)	singular common noun (e.g. book, girl)
2.	JJ	53,292	(7.0%)	general adjective (e.g. digital, international)
3.	AT	51,937	(6.8%)	article (e.g. the, no)
4.	II	39,516	(5.2%)	general preposition (e.g. to, in)
5.	NN2	37,305	(4.9%)	plural common noun (e.g. books, girls)
Σ		353,005	(46.2%)	

Table 3: Distribution of number of tags per token in running text for the English part of the CRATER corpus.

# tags	Frequency		# tags	Frequency	
1	204,817	(26.78%)	8	29,190	(3.82%)
2	159,573	(20.87%)	10	28,413	(3.72%)
3	132,411	(17.32%)	12	14,829	(1.94%)
4	81,163	(10.61%)	20	3,852	(0.50%)
5	35,628	(4.66%)	> 1	559,860	(73.22%)
6	63,140	(8.26%)			
7	11,661	(1.52%)			

Average: 3.47 tags/token

Table 4: 5 most ambiguous words in the English part of the CRATER corpus.

	word	#tags	frequency	
1.	as	20	3,852	(0.5%)
2.	to	12	14,538	(1.9%)
3.	so	12	291	(0.04%)
4.	a	10	14,948	(2.0%)
5.	in	10	13,465	(1.8%)

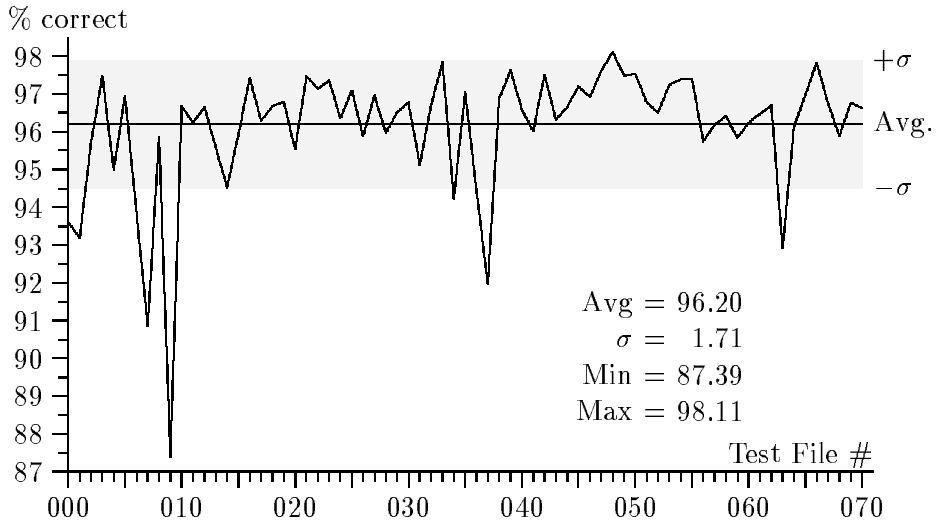

Figure 1: Percentage of correctly tagged words for the English part of the CRATER corpus with statistical trigram tagging. One file is used for testing, all other files for training. Unknown words are handled by analyzing the suffix of length 3, sparse data is handled by estimated likelihood estimation (addition of 0.5).

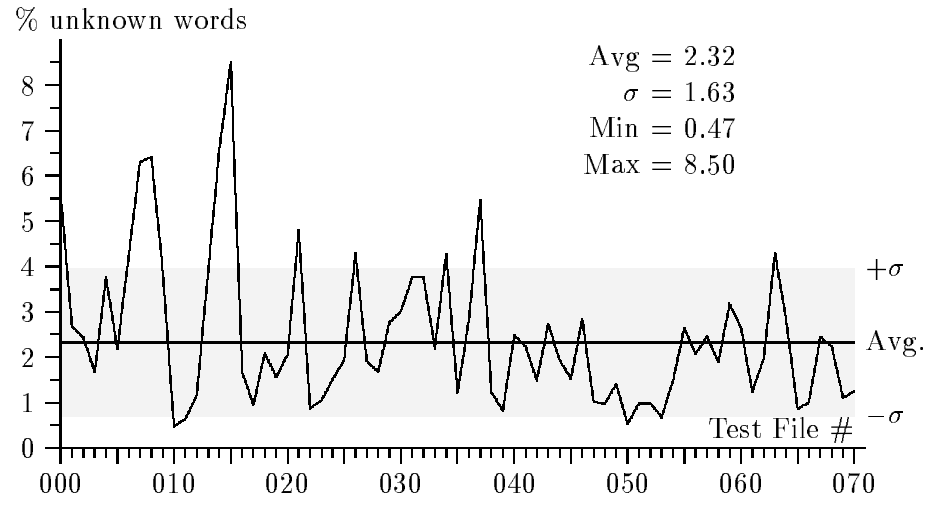

Figure 2: Percentage of unknown words in each file of the English part of the CRATER corpus.

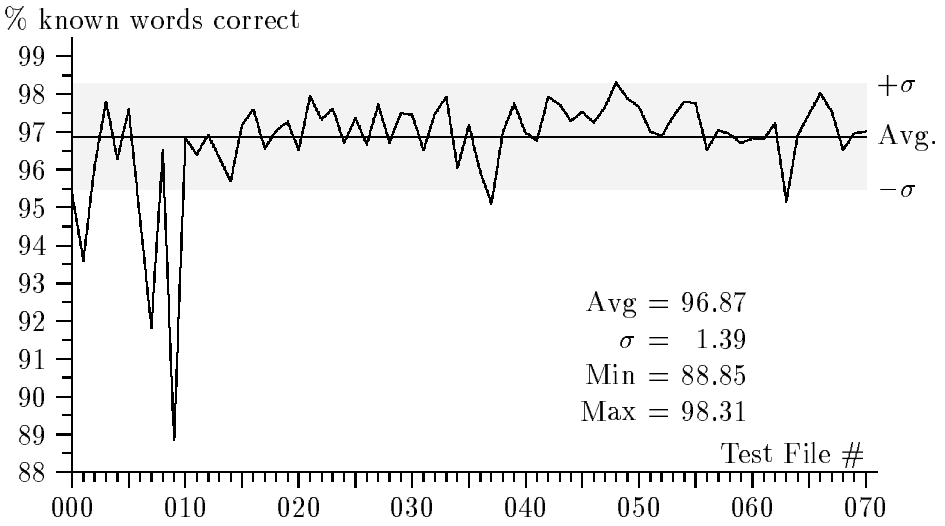

Figure 3: Percentage of correctly tagged known words (i.e., words that also occured in the training part) for the English part with statistical trigram tagging. One file is used for testing, all other files for training.

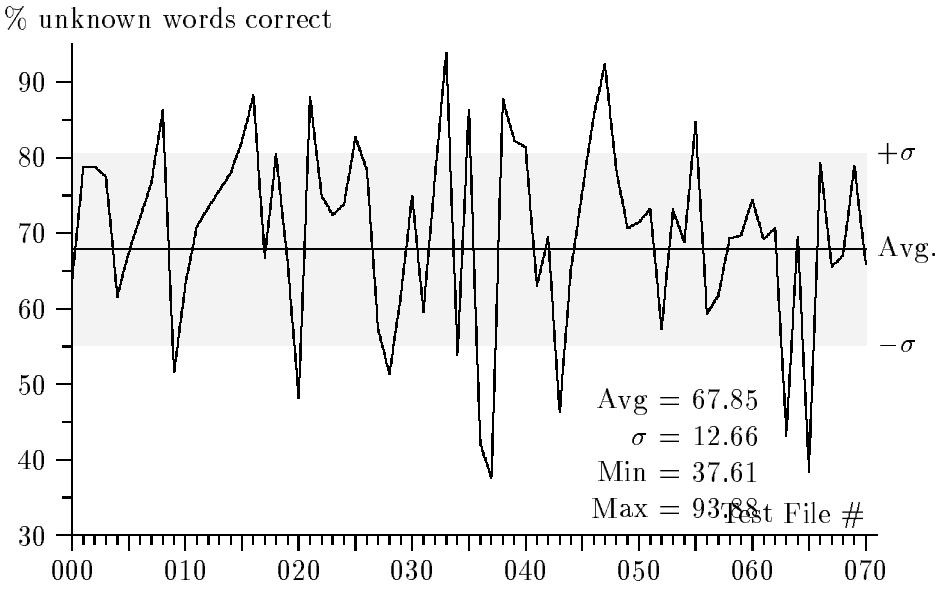

Figure 4: Percentage of correctly tagged unknown words (i.e., words that did not occur in the training part) for the English part with statistical trigram tagging. One file is used for testing, all other files for training.

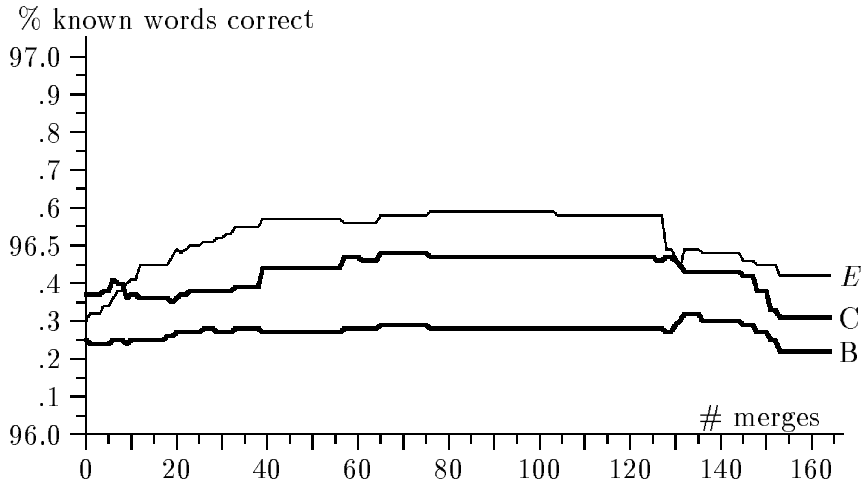

Figure 5: Tagset clustering for the English CLAWS tagset. The diagram shows the tagging results after each step of merging, starting with 217 tags, ending after 164 merges with 53 tags. All information of the original tagset is preserved.

Table 5: Tags involved in the first 10 merges of the CLAWS tagset, reducing the tagset without information loss. $[n]$ denotes the cluster generated in step n.

	tag 1	tag 2
1.	NNU	NNL2
2.	[1]	NPM1
3.	[2]	NN121
4.	RL	REX42
5.	[3]	NNJ2

	tag 1	tag 2
6.	CS32	CSW32
7.	MC	MD
8.	[5]	NNl1
9.	CC	CS
10.	RA	RPK

Table 6: 10 most frequent tokens in the Spanish part of the CRATER corpus, not distinguishing upper and lower case letters.

	Token	Frequency			Token	Frequency	
1.	de	39,958	(8.4%)	6.	en	10,398	(2.2%)
2.	la	18,190	(3.8%)	7.	se	8,444	(1.8%)
3.	,	15,032	(3.2%)	8.	)	8,239	(1.7%)
4.	.	14,489	(3.1%)	9.	a	6,587	(1.4%)
5.	el	12,640	(2.7%)	10.	que	6,429	(1.4%)
				Σ		140,406	(29.6%)

Table 7: 5 most frequent tags used in the Spanish part of the CRATER corpus.

	Tag	Frequency		Description
1.	PREP	69,038	(14.6%)	Preposition (a, con)
2.	NCFS	45,752	(9.7%)	Feminine singular common noun (mesa, mano)
3.	NCMS	43,090	(9.1%)	Masculine singular common noun (libro, ordenador)
4.	ARTDFS	18,086	(3.8%)	Feminine singular definite article (la)
5.	CM	15,032	(3.2%)	Comma (,)
Σ		190,998	(40.3%)	

Table 8: Distribution of number of tags per token in running text for the Spanish part of the CRATER corpus.

# tags	frequency		# tags	frequency	
1	264,922	(55.91%)	6	191	(0.04%)
2	99,427	(20.98%)	8	495	(0.10%)
3	76,222	(16.09%)	> 1	208,925	(44.09%)
4	31,873	(6.73%)			
5	717	(0.15%)	Average: 1.75 tags/token		

Table 9: 5 most ambiguous words in the Spanish part of the CRATER corpus.

	word	#tags	frequency	
1.	s	8	495	(0.10%)
2.	base	6	174	(0.04%)
3.	dependiente	6	17	(0.004%)
4.	local	5	179	(0.04%)
5.	correspondientes	5	111	(0.02%)

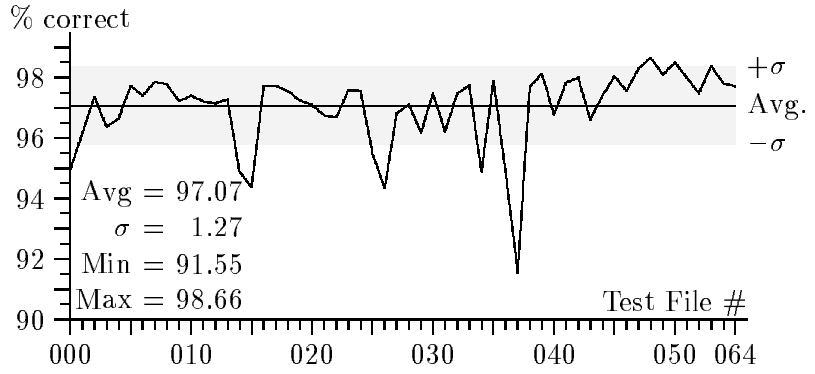

Figure 6: Percentage of correctly tagged words with statistical trigram tagging for the Spanish part of the CRATER corpus. One file is used for testing, all other files for training. Unknown words are handled by analyzing the suffix of length 3, sparse data is handled by estimated likelihood estimation (addition of 0.5).

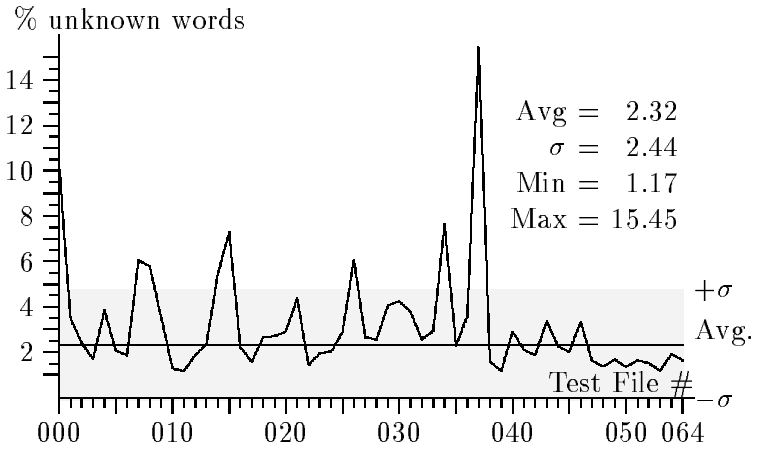

Figure 7: Percentage of unknown words in each file of the Spanish part of the CRATER corpus.

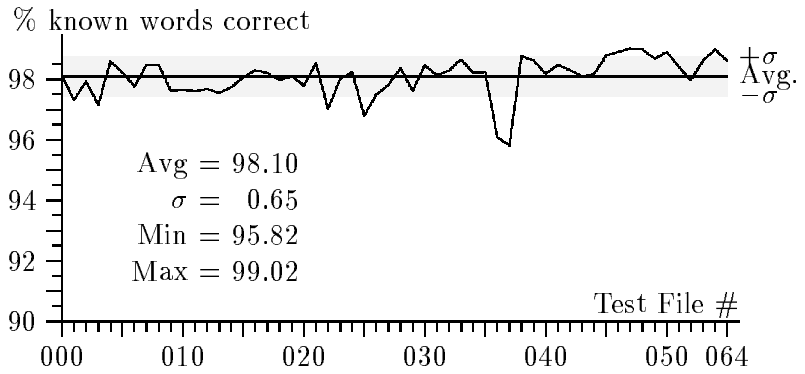

Figure 8: Percentage of correctly tagged known words (i.e., words that also occured in the training part) for the Spanish part with statistical trigram tagging. One file is used for testing, all other files for training.

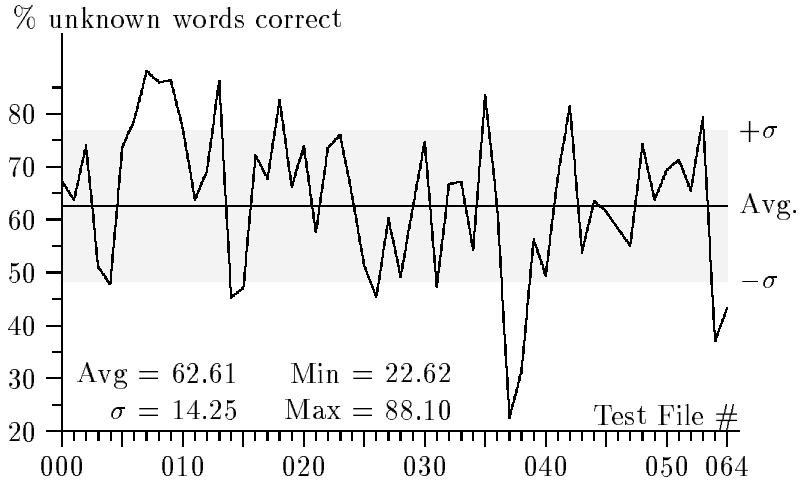

Figure 9: Percentage of correctly tagged unknown words (i.e., words that did not occur in the training part) for the Spanish part with statistical trigram tagging. One file is used for testing, all other files for training.