Automatic Stochastic Tagging of Natural Language Texts

Evangelos Dermatas*
University of Patras

George Kokkinakis*
University of Patras

Five language and tagset independent stochastic taggers, handling morphological and contextual information, are presented and tested in corpora of seven European languages (Dutch, English, French, German, Greek, Italian and Spanish), using two sets of grammatical tags; a small set containing the eleven main grammatical classes and a large set of grammatical categories common to all languages. The unknown words are tagged using an experimentally proven stochastic hypothesis that links the stochastic behavior of the unknown words with that of the less probable known words. A fully automatic training and tagging program has been implemented on an IBM PC-compatible 80386-based computer. Measurements of error rate, time response, and memory requirements have shown that the taggers' performance is satisfactory, even though a small training text is available. The error rate is improved when new texts are used to update the stochastic model parameters.

1. Introduction

In the natural language processing community, there has been a growing awareness of the key importance that lexical and corpora resources, especially annotated corpora, have to play, both in the advancement of research in this area and in the development of relevant products. In order to reduce the huge cost of manually creating such corpora, the development of automatic taggers is of paramount importance. In this respect, the ability of a tagger to handle both known and unknown words, to improve its performance by training, and to achieve a high rate of correctly tagged words, is the criterion for assessing its usability in practical cases.

Several taggers based on rules, stochastic models, neural networks, and hybrid systems have already been presented for Part-of-speech (POS) tagging. Rule-based taggers (Brill 1992; Elenius 1990; Jacobs and Zernik 1988; Karlsson 1990; Karlsson et al. 1991; Voutilainen, Heikkila, and Antitila 1992; Voutilainen and Tapanainen 1993) use POS-dependent constraints defined by experienced linguists. A small error rate has been achieved by such systems when a restricted, application-dependent POS set is used; e.g., an error rate of 2–6 percent has been reported by Marcus, Santorini, and Marcinkiewicz (1993) using the Penn Treebank corpus. Nevertheless, if a large POS set is specified, the number of rules increases significantly and rule definition becomes highly costly and cumbersome.

Stochastic taggers use both contextual and morphological information, and the model parameters are usually defined or updated automatically from tagged texts (Cerf-Danon and El-Beze 1991; Church 1988; Cutting et al. 1992; Dermatas and Kokkinakis 1988, 1990, 1993, 1994; Garside, Leech, and Sampson 1987; Kupiec 1992; Maltese

* Department of Electrical Engineering, Wire Communications Laboratory (WCL), University of Patras, 265 00 Patras, Greece. E-mail: dermatas@wcl.ee.upatras.gr.

and Mancini 1991; Meteer, Schwartz, and Weischedel 1991; Merialdo 1991; Pelillo, Moro, and Refice 1992; Weischedel et al. 1993; Wothke et al. 1993). These taggers are preferred when tagged texts are available for training, and large tagsets and multilingual applications are involved. In the case where additionally raw untagged text is available, the Maximum Likelihood training can be used to reestimate the parameters of HMM taggers (Merialdo 1994).

Connectionist models have been used successfully for lexical acquisition (Eineborg and Gamback 1993; Elenius 1990; Elenius and Carlson 1989; Nakamura et al. 1990). Correct classification rates up to 96.4 percent have been achieved in the latter case by testing on the Teleman Swedish corpus. On the other hand, a time-consuming training process has been reported.

Recently, several solutions to the problem of tagging unknown words have been presented (Charniak et al. 1993; Meteer, Schwartz, and Weischedel 1991). Hypotheses for unknown words, both stochastic (Dermatas and Kokkinakis 1993, 1994; Maltese and Mancini 1991; Weischedel et al. 1993), and connectionist (Eineborg and Gamback 1993; Elenius 1990) have been applied to unlimited vocabulary taggers. In taggers that are based on hidden Markov models (HMM), parameters of the unknown words are estimated by taking into account morphological information from the last part of the word (Dermatas and Kokkinakis 1994; Maltese and Mancini 1991). Accurate tagging of seven European languages has been achieved in the first case (error rates of 3–13 percent for a detailed POS set), but an enormous amount of training text is required for the estimation of the parameters for unknown words. Similar results have been reported by Maltese and Mancini (1991) for the Italian language. Weischedel et al. (1993) have used four categories of word morphology, such as inflectional endings, derivational endings, hyphenation, and capitalization. For the case in which only a restricted training text is available, a simple, language- and tagset-independent HMM tagger has been presented by Dermatas and Kokkinakis (1993), where the HMM parameters for the unknown words are estimated by assuming that the POS probability distribution of the unknown words and the POS probability distribution of the less probable words in the small training text are identical.

In this paper, five natural language stochastic taggers that are able to predict POS of unknown words are presented and tested following the process of developing annotated corpora (the most recently fully tagged and corrected text is used to update the model parameters). Three stochastic optimization criteria and seven European languages (Dutch, English, French, German, Greek, Italian and Spanish) and two POS sets are used in the tests. The set of main grammatical classes and an extended set of detailed grammatical categories is the same in all languages. The testing material consists of newspaper texts with 60,000–180,000 words for each language and an English EEC-law text with 110,000 words. This material was assembled and annotated in the framework of the ESPRIT-291/860 project "Linguistic Analysis of the European Languages." In addition, we present transformations of the taggers' calculations to a fixed-point arithmetic system, which are useful for machines without floating-point hardware.

The taggers handle both lexical and tag transition information, and without performing morphological analysis can be used to annotate corpora when small training texts are available. Thus, they are preferred when a new language or a new tagset is used. When the training text is adequate to estimate the tagger parameters, more efficient stochastic taggers (Dermatas and Kokkinakis 1994; Maltese and Mancini 1991; Weischedel et al. 1993) and training methods can be implemented (Merialdo 1994).

The structure of this paper is as follows: in Section 2 the stochastic tagging models are presented in detail. In Section 3 the influence of the training text errors and the sources of stochastic tagger errors are discussed, followed, in Section 4, by a short presentation of the implementation. In Section 5, statistical measurements on the corpora and a short description of the taggers' performance is given. Detailed experimental results are included in Appendices A and B.

2. Stochastic Tagging Models

A stochastic optimal sequence of tags T, to be assigned to the words of a sentence W, can be expressed as a function of both lexical $P(W \mid T)$ and language model $P(T)$ probabilities using Bayes' rule:

$$T_o = \underset{T}{\operatorname{argmax}} P(T \mid W) = \underset{T}{\operatorname{argmax}} \frac{P(W \mid T) * P(T)}{P(W)} = \underset{T}{\operatorname{argmax}} P(W \mid T) * P(T) \qquad (1)$$

Several assumptions and approximations on the probabilities $P(W \mid T)$ and $P(T)$ lead to good comprises concerning memory and computational complexity.

2.1 Hidden Markov Model (HMM) Approach

The tagging process can be modeled by an HMM by assuming that each hidden tag state produces a word in the sentence, each word w_i is uncorrelated with neighboring words and their tags, and each tag is probabilistic dependent on the N previous tags only.

2.1.1 Most probable tag sequence (HMM-TS). The optimal tag sequence for a given observation sequence of words is given by the following equation:

$$T_o^{(\mathrm{HMM}-TS)} = \underset{t_1,\ldots,t_M}{\operatorname{argmax}} P(t_1) \prod_{i=2}^{N} P(t_i \mid t_{i-1},\ldots,t_1) \prod_{i=N+1}^{M} P(t_i \mid t_{i-1},\ldots,t_{i-N}) \prod_{i=1}^{M} P(w_i \mid t_i) \qquad (2)$$

where M is the number of words in the sentence W.

The optimal solution is estimated by the well-known Viterbi algorithm. The first- (Rabiner 1989) and second- (He 1988) order Viterbi algorithms have been presented elsewhere. Recently, Tao (1992) described the Viterbi algorithm for generalized HMMs.

2.1.2 Most probable tags (HMM-T). The optimal criterion is to choose the tags that are most likely to be computed independently at each word event:

$$T_o^{(\mathrm{HMM}-T)} = \{t_{io}, t_{io} = \underset{t_i}{\operatorname{argmax}} P(t_i \mid W)\}, \qquad i = 1, M \qquad (3)$$

The optimum tag t_{io} is estimated using the probabilities of the forward-backward algorithm (Rabiner 1989):

$$t_{io} = \underset{t_i}{\operatorname{argmax}} P(t_i, W) = \underset{t_i}{\operatorname{argmax}} P(t_i, w_1, \ldots, w_i) P(w_{i+1}, \ldots, w_M \mid t_i) \qquad (4)$$

The probabilities in equation 4 are estimated recursively for the first- (Rabiner 1989) and second-order HMM (Watson and Chung 1992).

The main difference between the optimization criteria in 2.1.1 and that in 2.1.2 results from the definition of the expected correct tagging rate; the HMM-TS model maximizes the correctly tagged sentences, while the HMM-T model maximizes the correctly tagged words.

2.1.3 Stochastic hypothesis for the unknown words. When a new text is processed, some words are unknown to the tagger lexicon (i.e. they are not included in the training text). In this case, in order to use the forward-backward and the Viterbi algorithm we must estimate the unknown word's conditional probabilities $P(w \mid t)$. Methods for the estimation of these probabilities have already been proposed (e.g. the use of word endings morphology). Nevertheless, these methods fail if only a small training text is available because of the huge number of events not occurring in this text, such as pairs of tags and word endings. To address the above problem we have approximated the conditional probabilities of the unknown word tags by the conditional probabilities of the less probable word tags, i.e. tags of the words occurring only once. In the following we demonstrate experimentally that this approximation is valid and independent of the training text size.

Figures 1 and 2 show the probability distributions of the tags in the training text (known words) and that of the words occurring only once in this text for the English and French language, respectively. Furthermore, the tags' probability distribution of the words that are not included in the training text and are characterized as unknown words is shown. This distribution is measured in a different open testing text, i.e. a text that may include both known and unknown words. The measurements were carried out on newspaper text and split into two parts of the same size—the training and the open testing text. Each part contained 90,000 words for the English text and 50,000 words for the French text. In this experiment, a tagset comprising the main grammatical categories was used: Verb (Ver), Noun (Nou), Adjective (Adj), Adverb (Adv), Pronoun (Pro), Preposition (Pre), Article/Determiner (A-D), Conjunction (Con), Particle (Par), Interjection (Int), Miscellaneous (Mis; i.e., tags that cannot be classified in the previous categories).

This experiment has two significant results:

a. *The probability distribution of the tags of unknown words is significantly different from the distribution for known words, while it is very close to the probability distribution of the tags of the less probable known words* both in the English and French text.

b. *A number of closed and functional grammatical classes has very low probability for both unknown and words occurring only once,* e.g., the tags article, determiner, conjunction, pronoun, miscellaneous in English text, and article, determiner, conjunction, pronoun, interjection and miscellaneous in French text.

In the English text, verbs, adjectives and conjunctions are more frequent than in the French text. On the other hand, prepositions in the French text have a 0.05 greater probability, which is also the most significant difference between the distributions of the two languages. Prepositions in the words occurring only once and in unknown words are minimal in the English text, while in the French text one out of ten unknown words is a preposition. The text coverage by prepositions is 11.2 percent for the English and 16.2 percent for the French corpus. This difference increases significantly in the lexicon coverage: 0.47 percent for the English and 1.54 percent for the French lexicon.

In Figures 3 and 4, the results of chi-square tests that measure the difference between the probability distribution of the tags of the less probable words and that of the unknown words are shown. Various sizes of training text and two sets of grammatical categories, the main set (11 classes) and an extended set (described in detail in Section 5) were used.

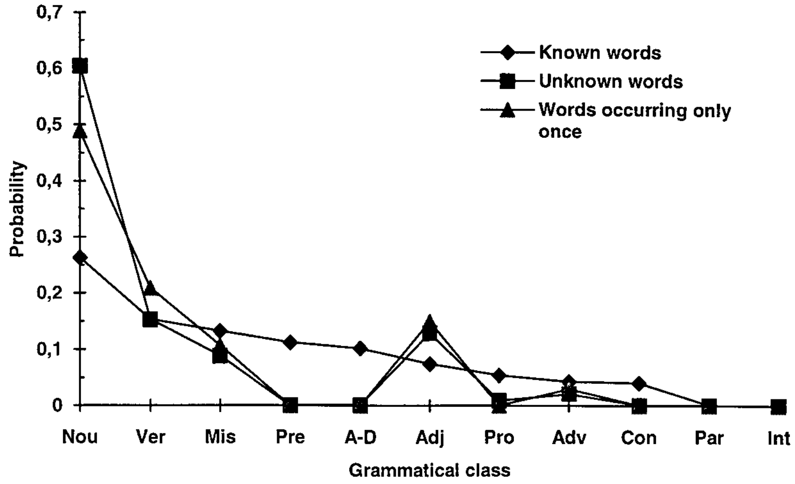


Figure 1
Distribution of the main grammatical classes of the known and unknown words and the words occurring only once in English text.

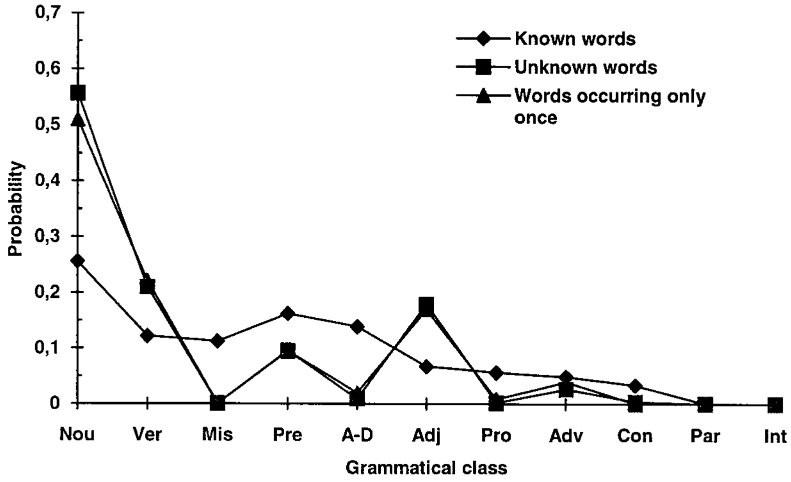


Figure 2
Distribution of the main grammatical classes of the known and unknown words and the words occurring only once in French text.

Specifically, the grammatically labeled text of 180,000 word entries of the English language was separated into two parts: the training text, where the tag probabilities distribution of the less probable words was estimated, and the open testing text, where the tag probabilities distribution of the unknown words was measured. Multiple chi-square experiments were carried out by transferring successively a portion of 30,000 words from the open testing text to the training text and by modifying the word occurrence threshold from 1 to 15 in order to determine the experimentally optimal threshold. Words having an occurrence below or equal to this threshold in the training text are counted as less probable words. The results of the tests shown in Figures 3 and 4 include threshold values up to 15 because the difference between the distributions for values greater than 15 increases significantly.

As shown in the above figures, the close relation between the tested probability distributions is evident for all sizes of training and testing text. Furthermore, we observe that:

a. The chi-square distance between the tag probability distributions is minimized for low values of the word occurrence threshold. In the tagset of main grammatical classes, this distance is minimized for threshold values less than three, four, or five, depending on the training text size. In the extended set of grammatical classes the distance is minimized in all cases for the threshold value one; i.e., when only the words occurring once in the training text are regarded as less probable words.

b. In the English text the chi-square distance between the tag probability distributions is minimized for 120,000 words training text for the set of main grammatical classes and for 60,000 words for the extended set. The same results are measured in the French text.

c. There is no significant variation in the chi-square test results for additional training text.

d. The closed and functional grammatical classes can be estimated automatically as the less probable grammatical classes of the less probable words in the tagged text. (The manual definition process is time-consuming when a set of detailed grammatical classes is used).

e. The probability distribution of some grammatical classes of the unknown words changes significantly when the size of the training text is increased. These changes can be measured in the training text from the tags' distribution of the less probable words.

Similar results have been achieved by testing the Dutch, German, Greek, Italian, and Spanish texts, both with the tagset of the main grammatical categories and with the common extended set of grammatical categories.

Based on the above we can complete both optimization criteria of the HMM formulation, given in 2.1.1 and 2.1.2, by calculating the conditional probability of the unknown word tags using Bayes' rule:

$$\begin{aligned} P(\text{Unknown word} \mid t_i) &= \frac{P(t_i \mid \text{Unknown word})P(\text{Unknown word})}{P(t_i)} \\ &\cong \frac{P(t_i \mid \text{Less probable word})P(\text{Unknown word})}{P(t_i)} \qquad (5) \end{aligned}$$

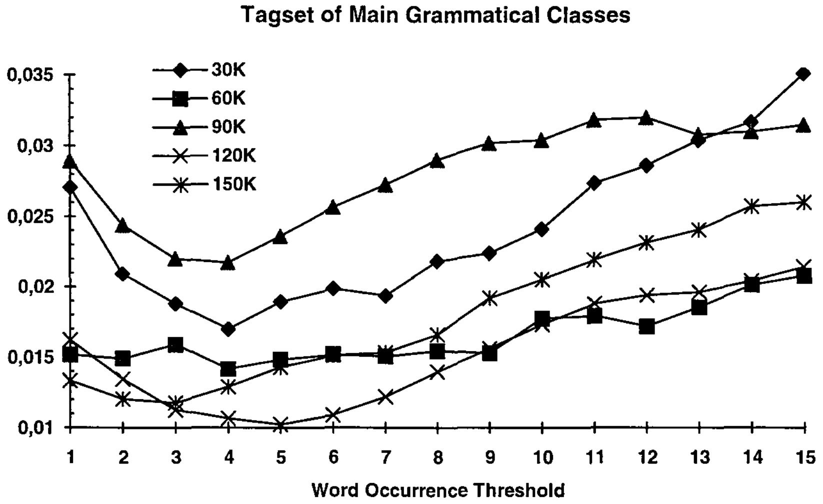


Figure 3
Chi-square test for the main grammatical classes' distribution of the unknown and the less probable words in the English text for various training text sizes.

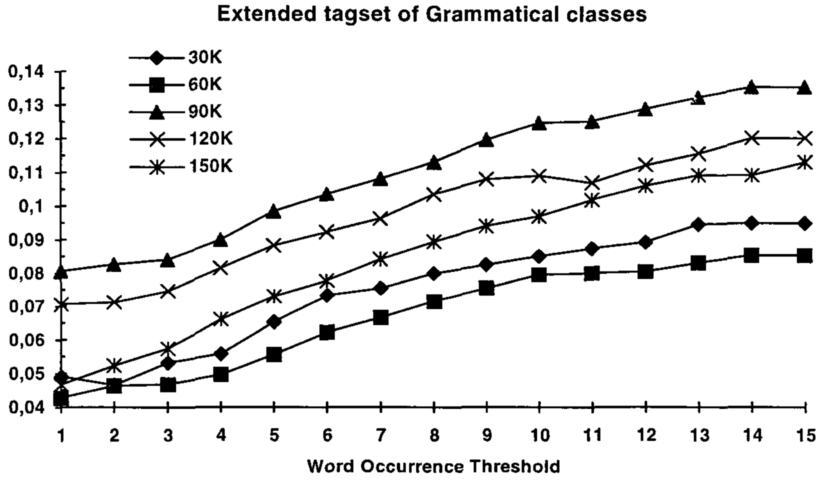


Figure 4
Chi-square test for the distribution of the grammatical tags of the unknown words and the less probable words in the English text, for the extended tagset of grammatical classes and various training text sizes.

The probability $P(\text{Unknown word})$ is approximated in *open testing* texts by measuring the unknown word frequency. Therefore *the model parameters are adapted each time an open testing text is being tagged.* The probability $P(t \mid \text{Less probable word})$ and the tags probability $P(t)$ are measured in the training text. Finally, each tag-conditional probability of the unknown word tags is normalized:

$$\sum_{j=1}^{L} P(w_j \mid t_i) + P(\text{Unknown word} \mid t_i) = 1, \qquad \forall i = 1, T \tag{6}$$

where L is the number of the known words and T is the number of tags.

2.2 Tagging without Lexical Probabilities

When the corresponding lexical probabilities $p(w \mid t)$ are not available in the dictionary that specifies the possible tags for each word, a simple tagger can be implemented by assuming that each word w_i in a sentence is uncorrelated with the assigned tag t_i; e.g., $p(w_i \mid t_i) = p(w_i)$.

In this case the most probable tag sequence, according to equation 2, is given by:

$$T_o^{(MLM)} = \underset{t_1,\ldots,t_M}{\operatorname{argmax}} P(t_1) \prod_{i=2}^{N} P(t_i \mid t_{i-1},\ldots,t_1) \prod_{i=N+1}^{M} P(t_i \mid t_{i-1},\ldots,t_{i-N}) \tag{7}$$

which is a Nth-order Markovian chain for the language model (MLM).

Taggers based on MLM require the training process to store each tag assigned to every lexicon entry and to define the unknown word tagset.

2.2.1 Stochastic hypothesis for the unknown words. The unknown word tagset is defined by the selection of the most probable tags that have been assigned to the less probable words of the training text. In this way the unknown words' ambiguity is decreased significantly. The word occurrence threshold used to define the less probable words and a tag probability threshold used to isolate the less probable tags are estimated experimentally.

Extensive experiments have shown insignificant differences in the tagging error rate when alternative word occurrence thresholds have been tested. The best results are obtained when values less than 10 are used. In this paper the word occurrence threshold has been set to one in all experiments.

3. Tagger Errors

3.1 Errors in the Training Text

Taggers based on the HMM technique compensate for some serious training problems inherent in the MLM approach. The most important one is the presence of errors in the training text. This situation appears when uncorrected tags or analysts' mistakes remain in the text used to estimate the stochastic model parameters. These errors generate tag assignments that are not valid. In MLM taggers these tags are equally weighted to the correct ones. In contrast, in HMM taggers invalid assignments are biased by the very low value of the corresponding conditional probability of the tags (the wrong tag rarely appears in the specific word environment), which decreases the overall probability for incorrect tag assignments.

Another important issue concerns the HMM ability to handle lexicon information, e.g., to find how frequently the tags have been assigned to each lexicon entry. In some languages, taggers based on HMMs almost reduce the prediction error to the half compared to the MLM approach.

3.2 Tagger prediction errors

Generally, tagger errors can be classified into three categories:

a. *Errors due to inadequate training data.* When the model parameters are estimated from a limited amount of training data, tagging errors appear because of unknown or inaccurately estimated conditional probabilities. Various interpolation techniques have been proposed for the estimation of the model parameters for unseen events or to smooth the model parameters (Church and Gale 1991; Essen and Steinbiss 1992; Jardino and Adda 1993; Katz 1987; McInnes 1992).

b. *Errors due to the syntactical or grammatical style of the testing text.* This type of error appears when the testing text has a style unknown to the model (i.e., a style used in the open testing text, not included in the training text). It can be reduced by using multiple models that have been previously trained in different text styles.

c. *Errors due to insufficient model hypotheses.* In this case the model hypotheses are not satisfied; e.g., there are strong intra-tag relations in distances greater than the model order, idiomatic expressions, language dependent exceptions, etc. A general solution to the variable length and depth of dependency for HMM has been already proposed (Tao 1992), but has not been implemented in taggers.

4. Implementation

In this section we present techniques to speed up the tagging process and avoid underflow or overflow phenomena during the estimation of the optimum solution. These techniques do not increase the prediction error rate or have only minimal influence on it, as proven in the experiments.

Two modules consume the majority of the tagger computational time. The first module extracts from the model parameters the intra-tag and the word-tag conditional probabilities requested by the second module, which computes the optimum solution by multiplying the corresponding conditional probabilities. Binary search maximizes the searching speed of the first module, while the following three transformation techniques decrease the computing time of the second module, avoid underflow or overflow phenomena, and use the faster and low-cost fixed-point arithmetic system.

4.1 Logarithmic Transformation

The stochastic solutions described by equations 2 and 7 are computed by multiplying several conditional probabilities. The floating-point multiplications of these probabilities are transformed into an equal number of floating-point additions, by computing the logarithm of the optimum criterion probability. This technique solves the underflow problem which arises when many small probabilities are multiplied, and accelerates the tagger response time.

4.2 Fixed-Point Transformation

The fixed-point transformation converts the floating-point logarithmic additions into an equal number of fixed-point additions. It is realized by the following quantization process:

$$I_x = \text{Round}\left[\frac{I_{\max}}{M_w \ln(P_{\min})}\left(\ln(P_{\min}) - \ln(P_x)\right)\right] \tag{8}$$

where: P_x is a conditional probability, $P_{\min}$ is the minimum conditional probability in the model parameter set, $I_{\max}$ is the maximum integer of the fixed-point arithmetic system, M_w is the maximum number of words in a sentence and Round$[\cdot]$ is a quantization function mapping real numbers into the nearest integer.

After the logarithmic and the fixed-point transformation, equations 2 and 7 become:

$$\begin{aligned} I_o^{(\text{HMM}-TS)} &= \underset{t_1,\ldots,t_M}{\operatorname{argmax}} I(t_1) + \sum_{i=2}^{N} I(t_i \mid t_{i-1},\ldots,t_1) \\ &\quad + \sum_{i=N+1}^{M} I(t_i \mid t_{i-1},\ldots,t_{i-N}) + \sum_{i=1}^{M} I(w_i \mid t_i) \end{aligned} \tag{9}$$

$$I_o^{(MLM)} = \underset{t_1,\ldots,t_M}{\operatorname{argmax}} I(t_1) + \sum_{i=2}^{N} I(t_i \mid t_{i-1},\ldots,t_1) + \sum_{i=N+1}^{M} I(t_i \mid t_{i-1},\ldots,t_{i-N}) \tag{10}$$

The quantization function approximates the computations, producing theoretically differing solutions. In practice the prediction error differences measured for all languages, taggers, and tagsets were less than 0.02 percent.

4.3 Scaling

The solution obtained by the forward-backward algorithm cannot be logarithmically transformed because of the presence of summations. It is well known that for HMMs the forward and backward probabilities tend exponentially to zero. The scaling process introduced in this case multiplies the forward and backward probabilities by a scaling factor at selective word events in order to keep the computations within the floating-point dynamic range of the computer (Rabiner 1989).

4.4 Hardware–Software

The taggers have been realized under MS-DOS using a 32-bit C compiler. The lexicon size is limited by the available RAM. A mean value of 35 bytes per word is allocated. The tagger speed exceeds the rate of 500 word/sec in a 80386 (33MHz) for all languages and tagsets in text with known words. A maximum memory requirement of 930Kb has been measured in the experiments described in this paper.

A set of symbols and keywords (a sentence separators set) and the maximum length of a sentence are the only manually defined parameters when the HMM taggers are applied.

In the MLM taggers, the word occurrence threshold that isolates the less probable words and the tag probability threshold used to reject the less probable tags from the unknown words tagset are the manually defined parameters.

The training process has been designed to estimate or update the model parameters from fully tagged text without any manual intervention. Therefore, frequency measurements are defined or updated as model parameters instead of conditional

Table 1
Size of the corpora.

Text	Dutch	English	French	German	Greek	Italian	Spanish
Newspaper	110,000	180,000	100,000	100,000	120,000	160,000	60,000
EEC-Law	—	110,000	—	—	—	—	—

Table 2
ESPRIT 291/860: Project partners.

Country	Partner
England	Acorn Computers Limited
France	Centre National de la Recherche Scientifique (CNRS), LIMSI Division
Germany	Ruhr - Universitaet Bochum, Lehrstuhl fur Allgemeine Elektrotechnik und Akustik
Greece	University of Patras, Wire Communications Laboratory (WCL), Speech and Language Group
Italy	Ing. C. Olivetti & C., S.p.A.
Italy	Centro Studi Applicazioni in Tecnologie Avanzate - CSATA
Netherlands	Katholieke Universiteit Nijmegen, Dienst A-Faculteiten
Spain	Universidad National de Educacion a Distancia (UNED), Madrid

probabilities that are computed afterwards by using the corresponding relative frequencies.

5. Performance of the Systems

5.1 Taggers

Five taggers have been realized and tested using bi-POS and tri-POS transition probabilities. Specifically, the first- and the second-order MLM (MLM1 and MLM2, respectively), the first- and the second-order HMM of the most probable tag sequence criterion (HMM-TS1 and HMM-TS2, respectively), and the first-order HMM of the most probable tag criterion (HMM-T1) have been realized.

5.2 Corpora

The tagger performance has been measured in extensive experiments carried out on corpora of seven languages, English, Dutch, German, French, Greek, Italian and Spanish, annotated according to detailed grammatical categories. In Table 1, the type and the size of these corpora is shown. They are part of corpora selected in the framework of the ESPRIT-I project 291/860: "Linguistic Analysis of the European Languages" (1985–1989) by the project partners (Table 2) and annotated by using semi-automatic taggers. Manual correction was performed by experienced, native analysts for each language separately. In all languages the entries were tagged as they appeared in the text. In the German corpus, for example, where multiple words are concatenated, the words were not separated.

5.3 Tagsets

Two sets of grammatical tags were isolated from a unified set of grammatical categories defined in the ESPRIT I project 291/860 (ESPRIT-860, Internal report, 1986):

Table 3
Extended set of grammatical categories.

Main grammatical categories	Detailed grammatical information
Adjective, Noun, Pronoun	Regular base comparative superlative interrogative person number case
Adverb	Regular base comparative superlative interrogative
Article, Determiner, Preposition	Person number case
Verb	Tense voice mood person number case

Table 4
Number of grammatical tags.

Text	Dutch	English	French	German	Greek	Italian	Spanish
Main set	9	News: 10, Law: 10	10	11	11	10	10
Extended set	50	News: 43, Low: 36	14	116	443	121	121

Table 5
Word ambiguity in the newspaper corpus.

Tagset	English	Dutch	German	French	Greek	Italian	Spanish
Main set	1.336	1.111	1.3	1.69	1.209	1.62	1.197
Extended set	1.417	1.291	1.878	1.705	1.855	1.729	1.25

a. A common tagset of 11 main grammatical categories for each language, as described in 2.1.3.

b. An extended set including common categorization of the grammatical information for all languages, as shown in Table 3. In some languages a number of grammatical categories is not applicable. The depth of grammatical analysis and the grammatical structure of each language produce a different number of POS tags. In Table 4 the number of POS tags used for each language and each set of grammatical categories is shown.

5.4 Corpus Ambiguity

The corpus ambiguity was measured by the mean number of possible tags for each word of the corpus for both sets of grammatical tags (Table 5). The most ambiguous texts are the French, Italian, and English in the tagset of main grammatical classes and the German, Greek, Italian, and French in the extended set of grammatical categories.

In Figure 5 the percent occurrence of unknown words in an open testing text of 10,000 words is shown versus the size of the training text.

The Italian and Greek corpora have the greatest number of unknown words followed by the Spanish corpus (for the available results with restricted training text).

Taking into account the word ambiguity in the training text (Table 5), the occurrence of unknown words in the open testing text (Figure 5), and the hypothesis that the unknown word tagset and the application tagset are the same, the ambiguity of the open testing corpus for both sets of grammatical categories was computed for a 50,000-word training corpus (Table 6).

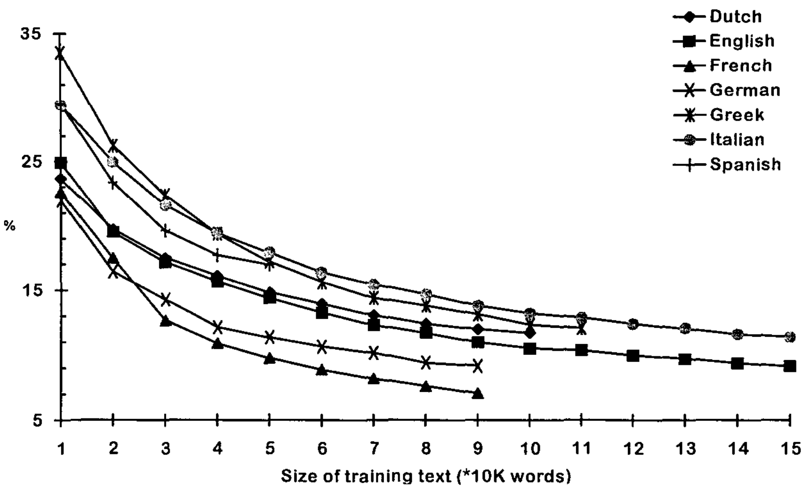


Figure 5
Percentage of unknown words in open testing text of 10,000 words for various sizes of the training text.

Table 6
Corpus ambiguity in newspaper open testing text.

Tagset	English	Dutch	German	French	Greek	Italian	Spanish
Main set	8.75	7.83	9.9	9.19	9.32	8.5	8.5
Extended set	37.03	42.78	103.07	12.8	367.25	99.86	100.69

For the set of main grammatical classes the ambiguity of the open testing corpus is more or less the same for all languages, varying from a minimum of 7.83 tags per word in the Dutch text to a maximum of 9.32 in the Greek corpus. For the extended set of grammatical categories three types of corpora can be distinguished:

a. The most ambiguous is the corpus of the Greek language, because of the great number of grammatical tags (443) and the strong presence of unknown words in the open testing text.

b. In the German, Spanish, and Italian texts the same ambiguity is measured.

c. The least ambiguous are the Dutch and French texts.

Taking into account the previous results, it is important to note that the great differences between languages in text ambiguity, in the presence of unknown words and in the statistics of the grammatical categories, e.g. the different occurrence of prepositions in English and French corpora, prevent a direct comparison of languages from the taggers' error rate. Apart from a few obvious observations given in Section 5.7, such a comparison would require a detailed examination of the corpora and the taggers' errors by experienced linguists. Therefore, the prediction error rates presented in

Table 7
Lexicon size for 100,000-word training text.

Language	Dutch	English	French	German	Greek	Italian
Lexicon size	13,700	12,200	13,500	8,900	17,400	15,300

this paper should be regarded only as indication of the probabilistic taggers' efficiency in each separate language when small training texts are available.

5.5 Experiments

The corpora were divided into 10,000-word entries. All parts except the last one were used to create (initially) and update the model parameters successively. The last part was tagged each time after the model parameters were updated, giving results of the tagger performance on open testing text. The influence of the application tagset on the tagger performance was measured by testing the two totally different tagsets described in Section 5.3.

The experimental process was repeated for each language, tagset and tagger. Thus a total number of 2 (tagsets) $*$ 5 (taggers) $*$ [7 (languages) $+1$ (Test on English EEC-law text)] $=80$ experiments was carried out.

5.6 Tagger Speed and Memory Requirements

In Figures 6 and 7 the tagger speed and the memory requirements after the last memory adaptation process are presented for all taggers and languages, and for the extended tagset.

The Greek and Italian corpora have a great number of lexical entries (different word forms) for the same amount of 100,000-word training text, as shown in Table 7. As a result these taggers require more memory (Figure 7). In contrast, the small size of the German lexicon decreases the required memory.

Tagger speed is closely related to the corpus ambiguity (Table 6). The ambiguity of the Greek corpus is more than three times greater than the next one, the German corpus.

The significant influence of the training text size on tagger speed is proven by comparing the experimental results in the English corpus (newspaper and EEC-Law). When the taggers are trained using the 170,000 words of the English newspaper corpus, a greater number of lexicon entries and a greater number of transition probabilities (Figure 7) is measured than in the case of the EEC-law corpus (100K words training text). The model becomes more complex, but tagger speed is slightly higher because of the greater size of the training text, which reduces the presence of unknown words in the testing text. Generally, tagger speed increases when the training text is increased.

5.7 Tagger Error Rate

The actual tagger error rates for all experiments are given in Appendices A and B. In this section we present a discussion of these error rates.

The error rate depends strongly on the test text and language, and the type and size of the tagset. The worst results have been obtained for the Greek language because of its significantly greater ambiguity, the number of tags (requiring significantly greater training text), and its freer syntax.

In the main category of tagset experiments, the model parameters for the MLM systems are estimated accurately when the training text exceeds 50,000–90,000 words,

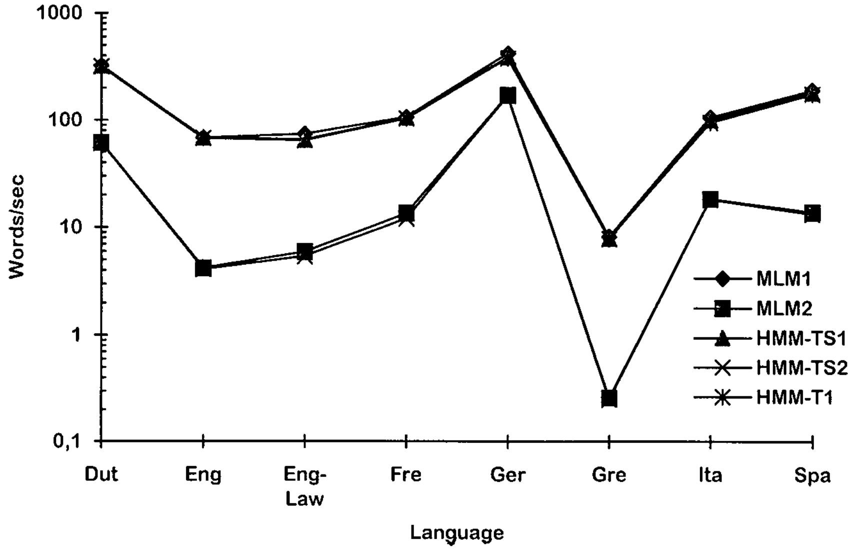


Figure 6
Tagger speed after the last adaptation process for the extended set of grammatical categories.

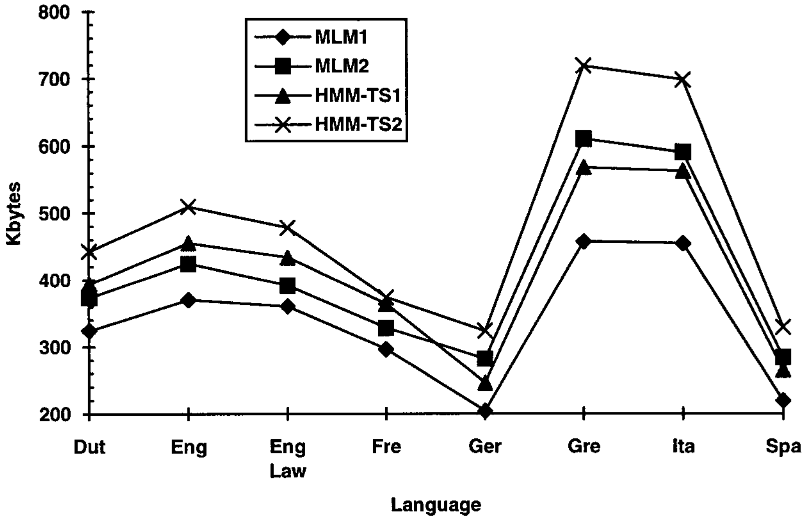


Figure 7
Tagger memory requirements for the extended set of grammatical categories.

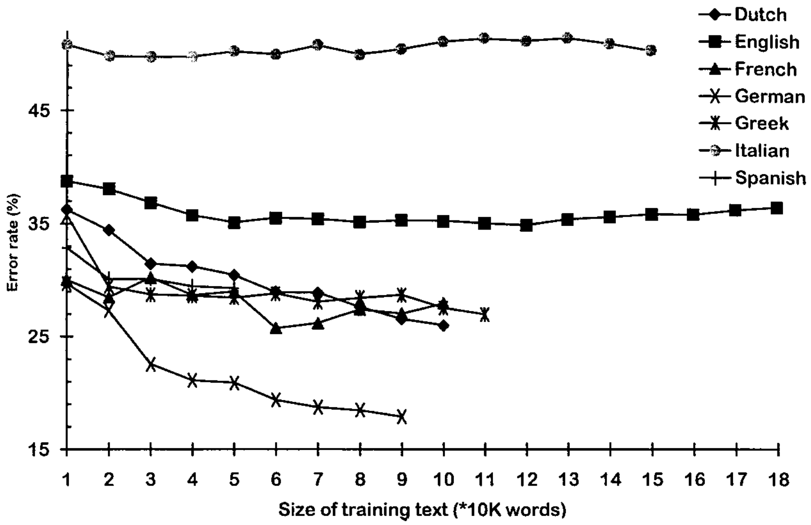


Figure 8
Unknown word error rate for the HMM-TS2 tagger and the set of main grammatical categories.

in contrast to the extended tagset experiments, where a greater-size training text for the German, Greek, and Spanish languages is required. This phenomenon becomes stronger in taggers based on the HMM where the accuracy of the $P(w \mid t)$ estimation is proportional to the word and the tag frequency of occurrence in the training text. Thus, for all tagsets and languages a larger training text is required in order to minimize the error rate.

The taggers based on the HMM reduce the prediction error almost to half in comparison to the same order taggers based on MLM. Strong dependencies on the language and the estimation accuracy of the model parameters influence this reduction. The alternative HMM solutions give trivial performance differences, confirming recent results obtained in the Treebank corpus by using an HMM tagger (Merialdo 1991).

Concerning the performance of the taggers in unknown words, we present in Figure 8 as an example the HMM-TS2 error rate for the tagset of the main grammatical categories, which is also the worst case for this set of grammatical categories. Generally the error rate decreases when the training text is increased. The stochastic model is successful for only half of the unknown words for the Italian text and for approximately two out of three unknown words for the English text. In all other languages the HMM-TS2 tagger gives the correct solution for three out of four unknown words.

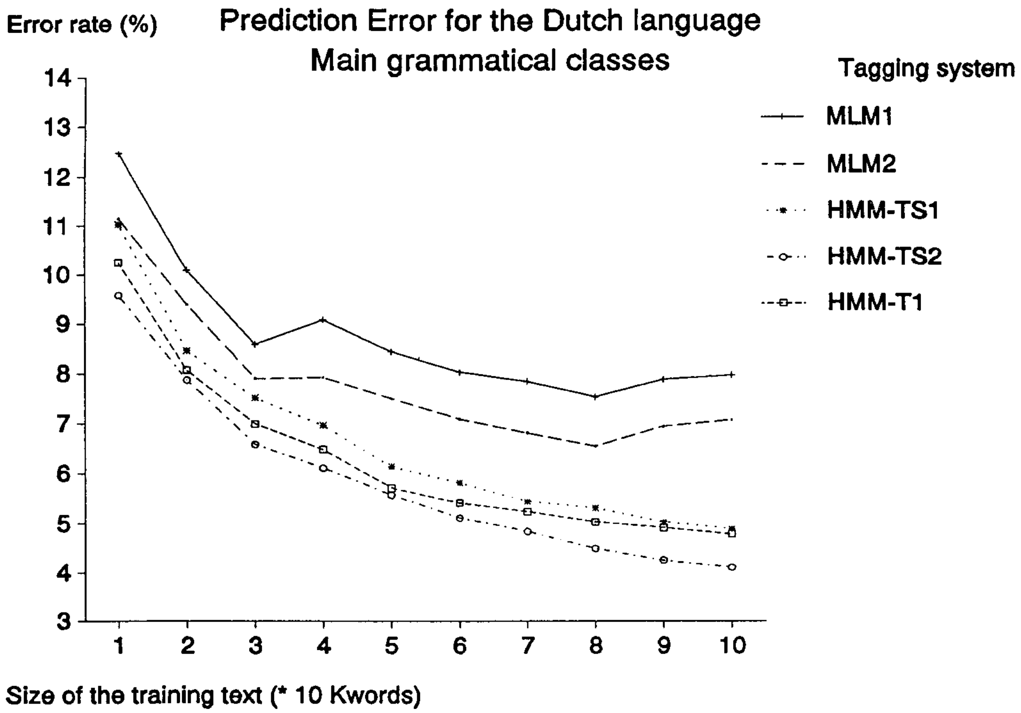
Similar results are achieved when the extended set of grammatical categories is tested. In this case the unknown word error rate increases about 10–20 percent for all the languages except the Greek language. In the Greek text the error rate reaches approximately 65 percent when 100,000-word text is used to define the parameters of the HMM.

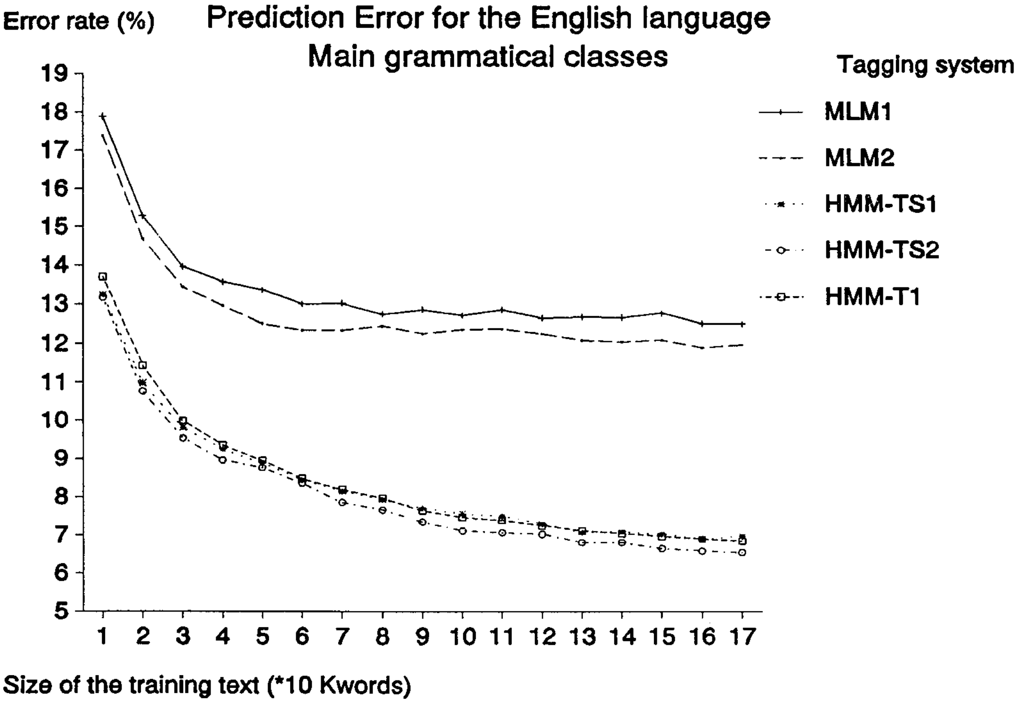
The unknown words, which initially cover about 25–35 percent of the text, are reduced to 8–15 percent when all the available text is used as training data. In the majority of the experiments, the tagger error rate decreases when new text updates the model parameters. Trivial differences of the tagger learning rates between languages and tagsets show the efficiency of the training method in estimating the model transition probabilities for the tested languages and the validity of the stochastic hypothesis for the unknown words.

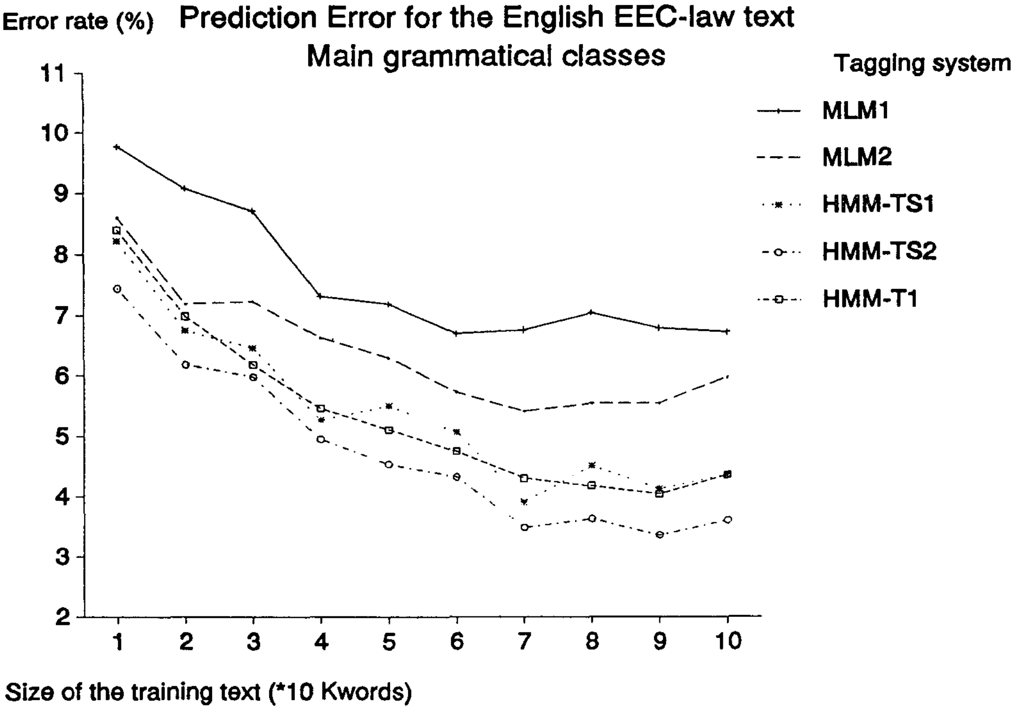
6. Conclusion

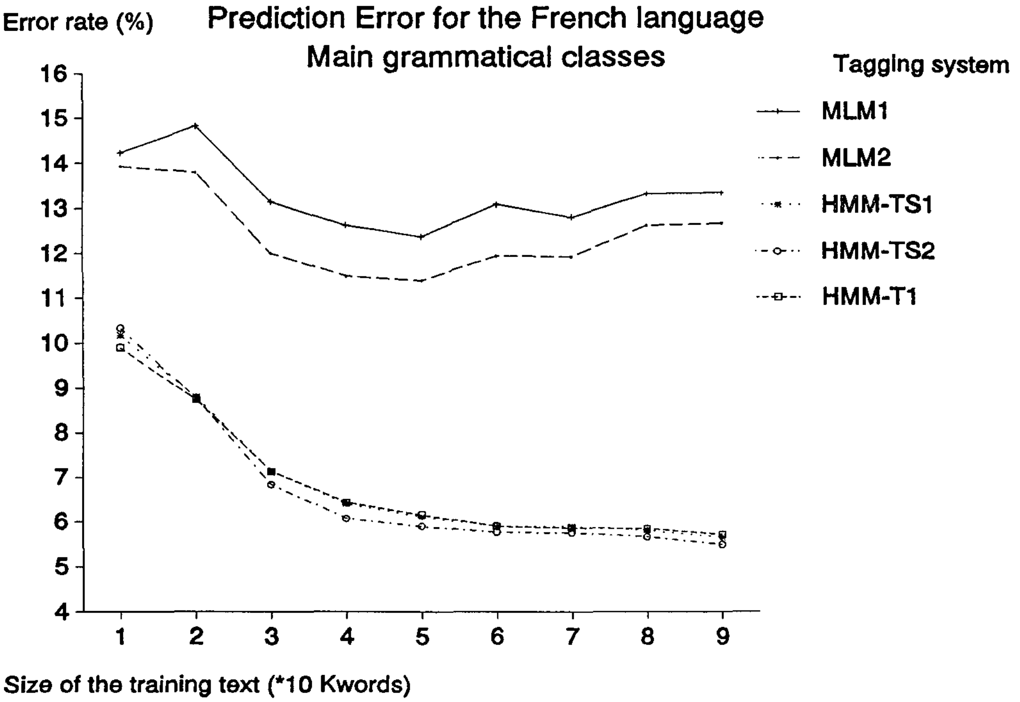
In this paper five automatic, stochastic taggers that are able to tag unknown words have been presented. The taggers have been tested in newspaper corpora of seven European languages and an EEC-law text of the English language using two sets of grammatical categories. When new training text updates the model parameters, the tagging error rate changes as expected: in text with unknown words a lower error rate is measured, proving the efficiency of the relative frequencies learning method and the validity of the hypothesis for the unknown words' stochastic behavior.

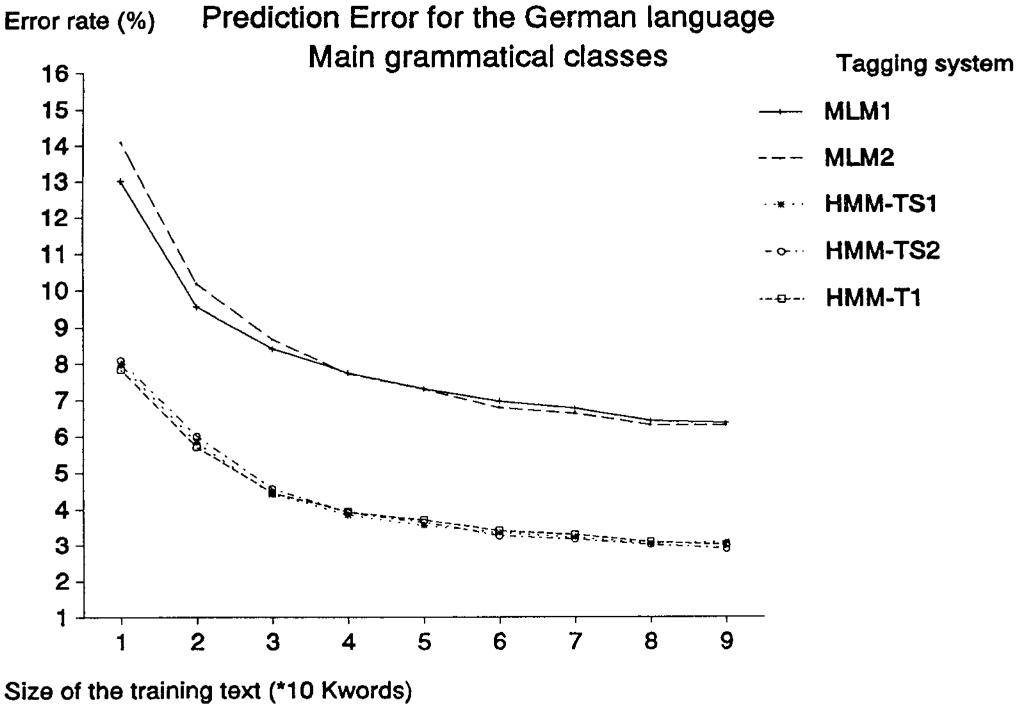
Appendix A: Tests in the Main Grammatical Categories Set

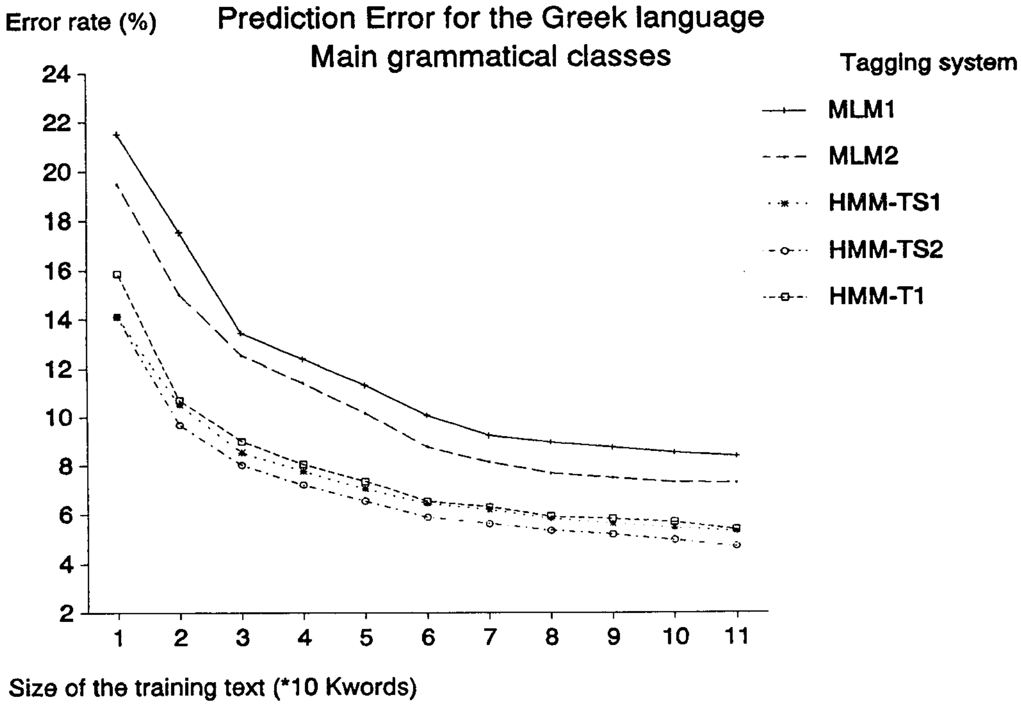


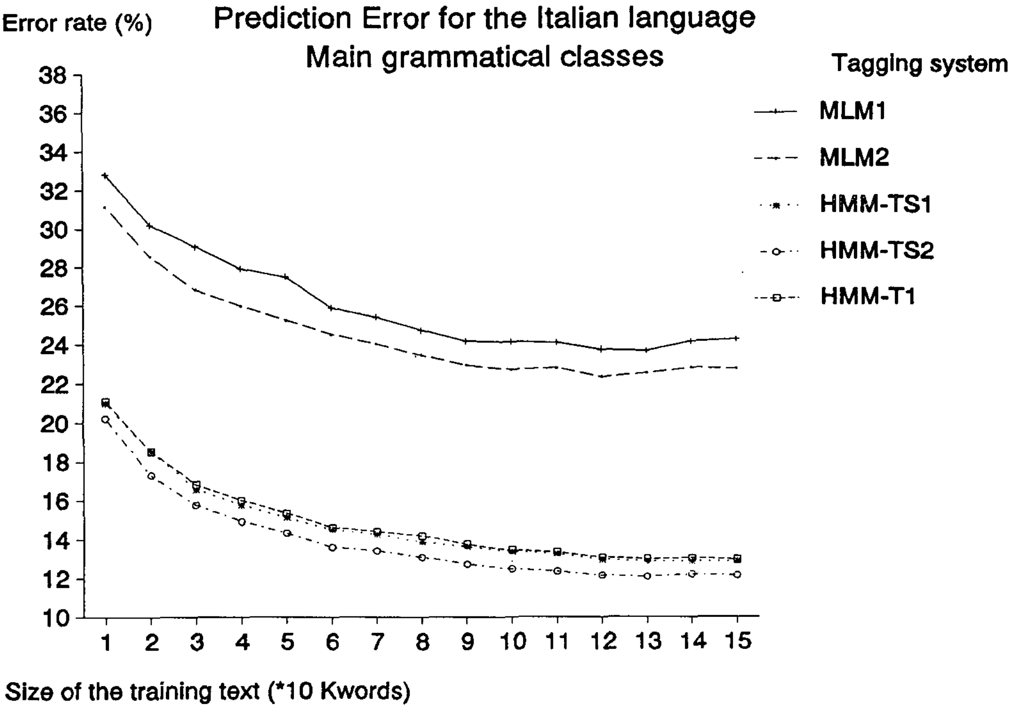


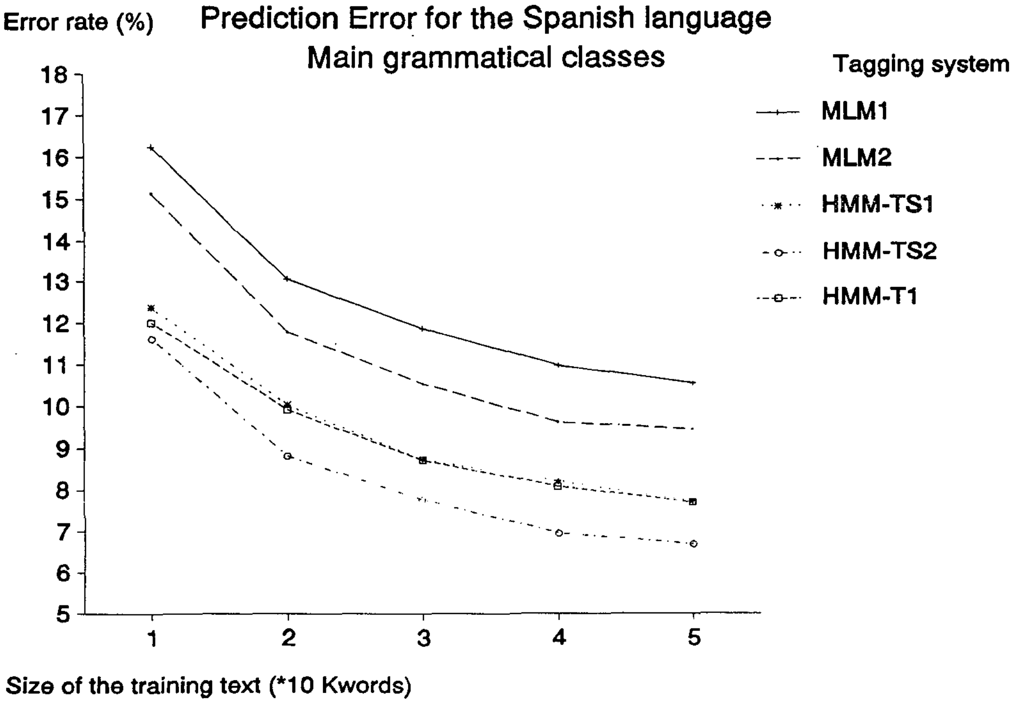

Error rate (%)
Prediction Error for the English EEC-law text
Main grammatical classes
Tagging system
MLM1
MLM2
HMM-TS1
HMM-TS2
HMM-T1
11
10
9
8
7
6
5
4
3
2
1
2
3
4
5
6
7
8
9
10
Size of the training text (*10 Kwords)


Error rate (%)
Prediction Error for the French language
Main grammatical classes
Tagging system
MLM1
MLM2
HMM-TS1
HMM-TS2
HMM-T1
16
15
14
13
12
11
10
9
8
7
6
5
4
1
2
3
4
5
6
7
8
9
Size of the training text (*10 Kwords)

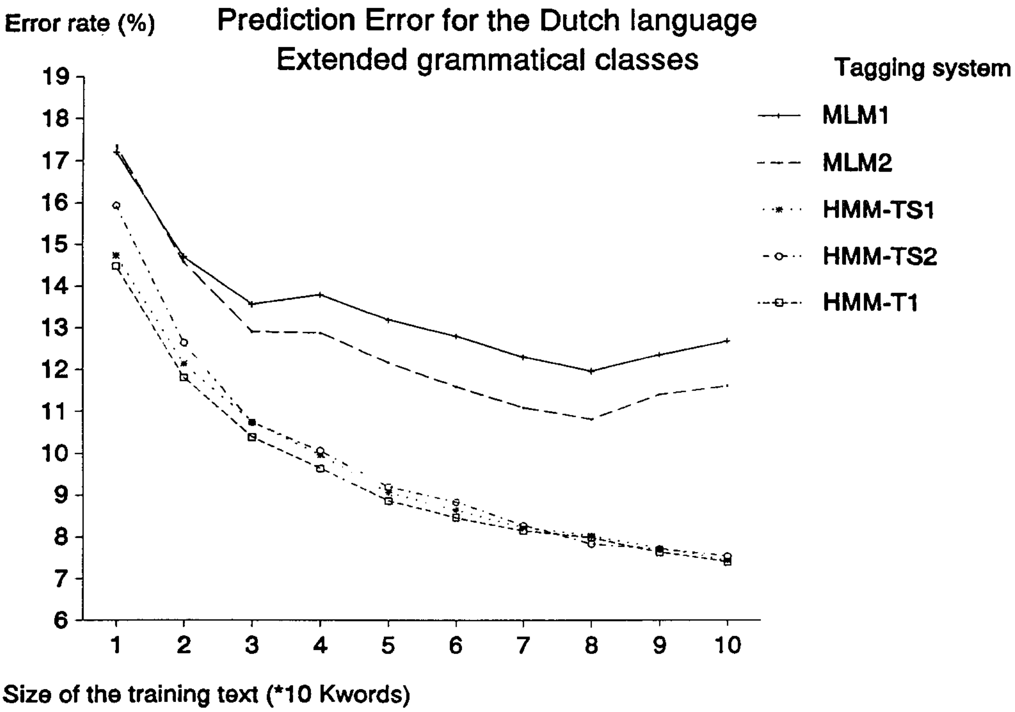

Error rate (%)
Prediction Error for the German language
Main grammatical classes
Tagging system
MLM1
MLM2
HMM-TS1
HMM-TS2
HMM-T1
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
Size of the training text (*10 Kwords)

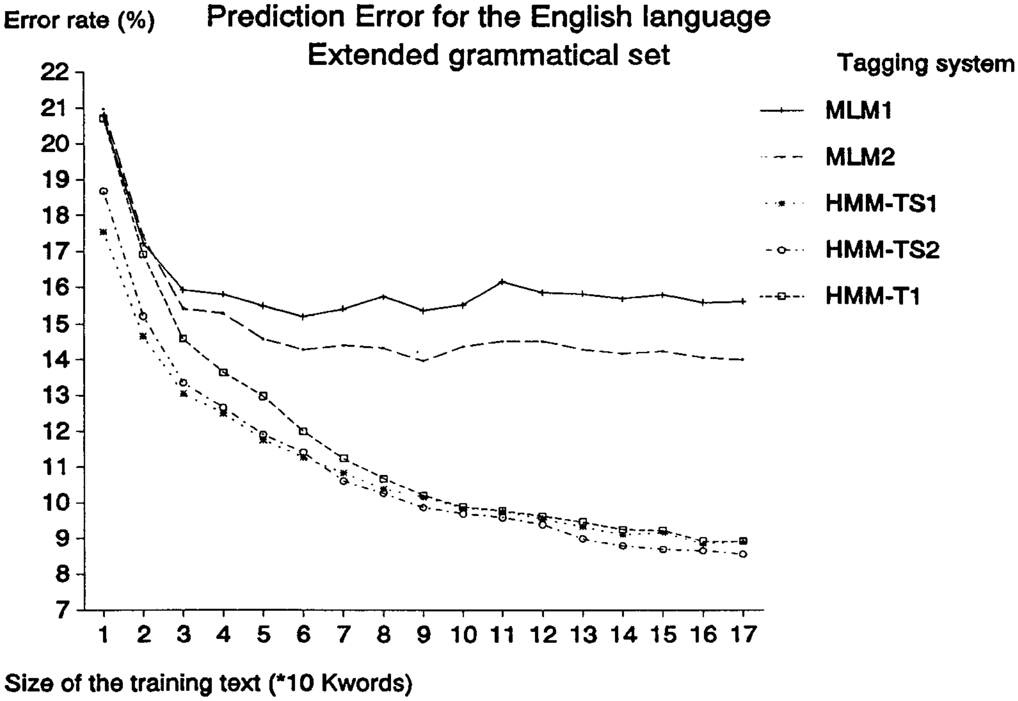

Error rate (%)
Prediction Error for the Greek language
Main grammatical classes
Tagging system
MLM1
MLM2
HMM-TS1
HMM-TS2
HMM-T1
24
22
20
18
16
14
12
10
8
6
4
2
1
2
3
4
5
6
7
8
9
10
11
Size of the training text (*10 Kwords)

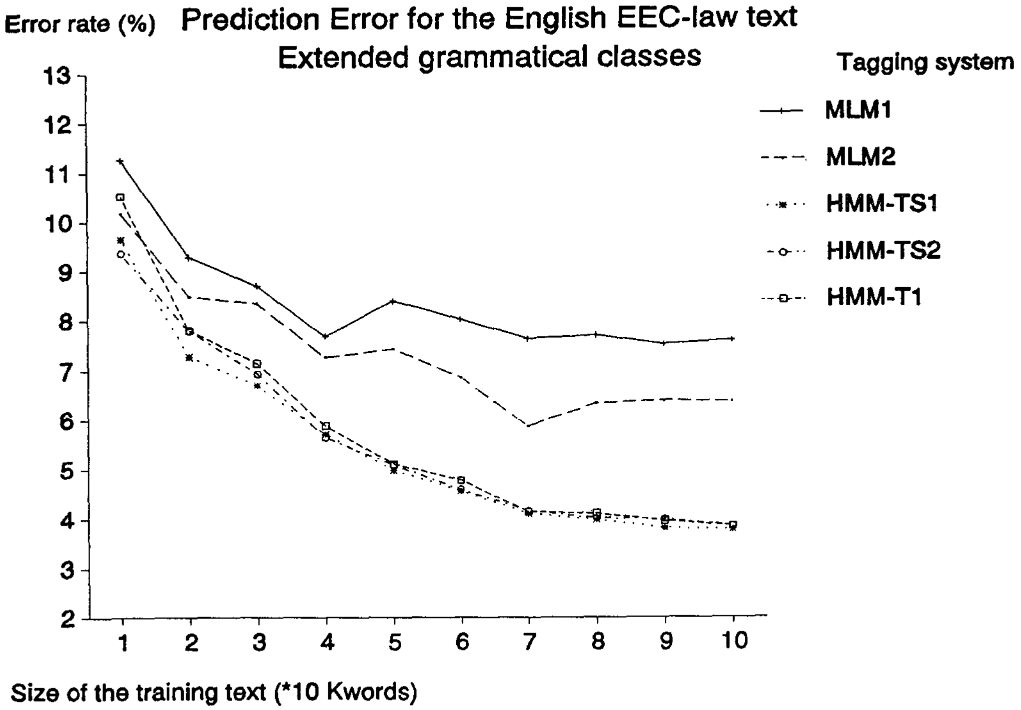

Error rate (%)
Prediction Error for the Italian language
Main grammatical classes
Tagging system
MLM1
MLM2
HMM-TS1
HMM-TS2
HMM-T1
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Size of the training text (*10 Kwords)

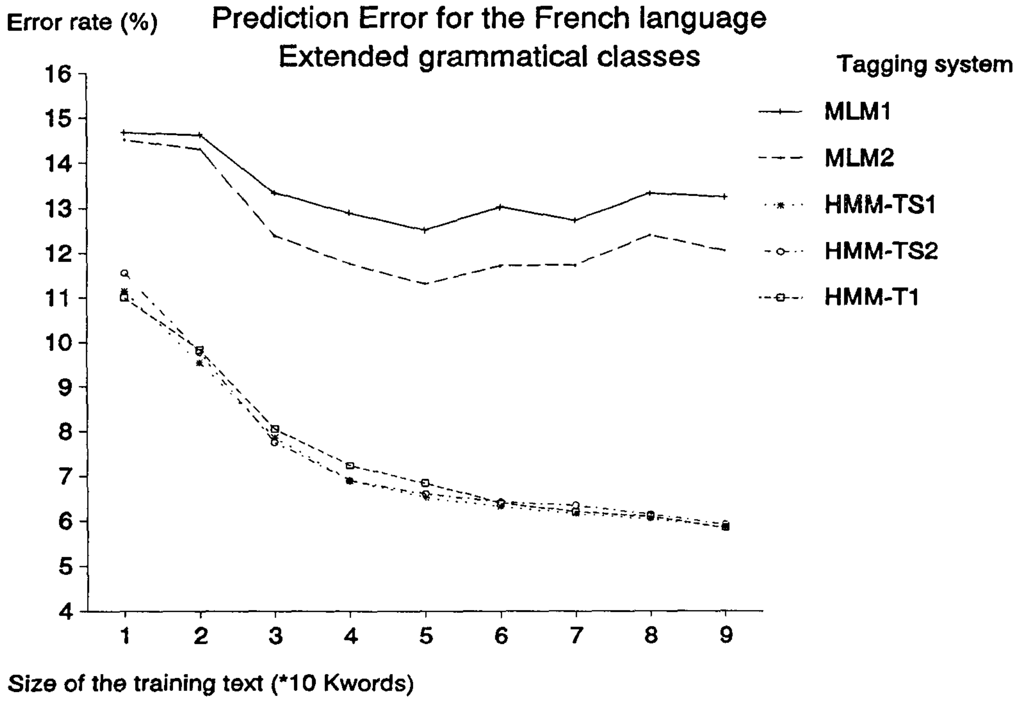

Error rate (%)
Prediction Error for the Spanish language
Main grammatical classes
Tagging system
MLM1
MLM2
HMM-TS1
HMM-TS2
HMM-T1
18
17
16
15
14
13
12
11
10
9
8
7
6
5
1
2
3
4
5
Size of the training text (*10 Kwords)

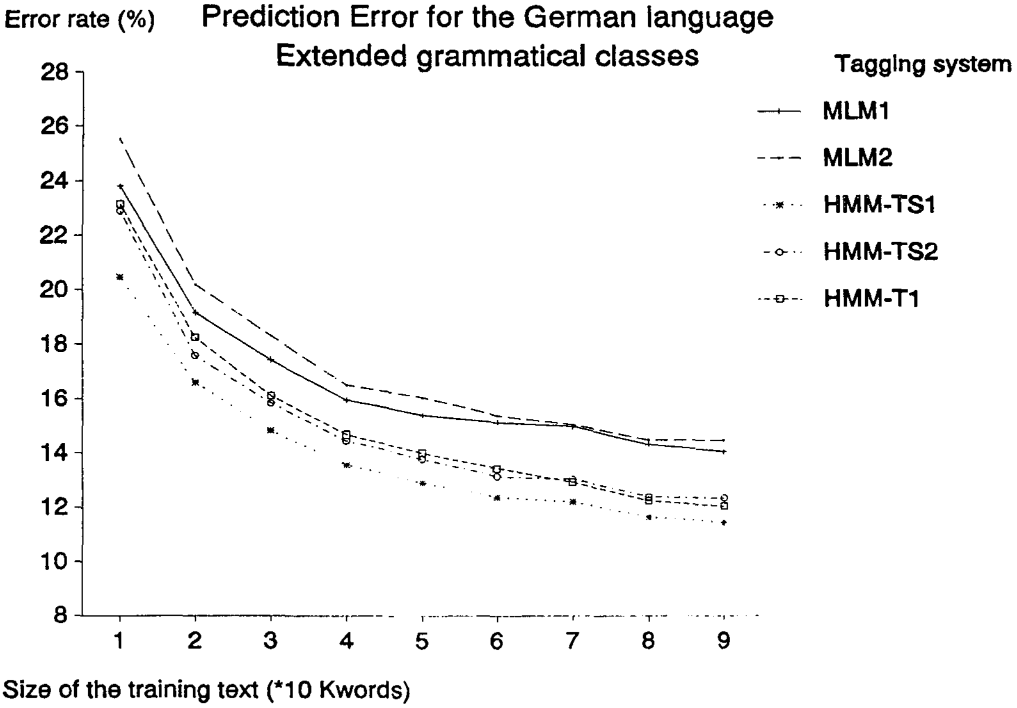
Appendix B: Tests in the Extended Grammatical Categories Set

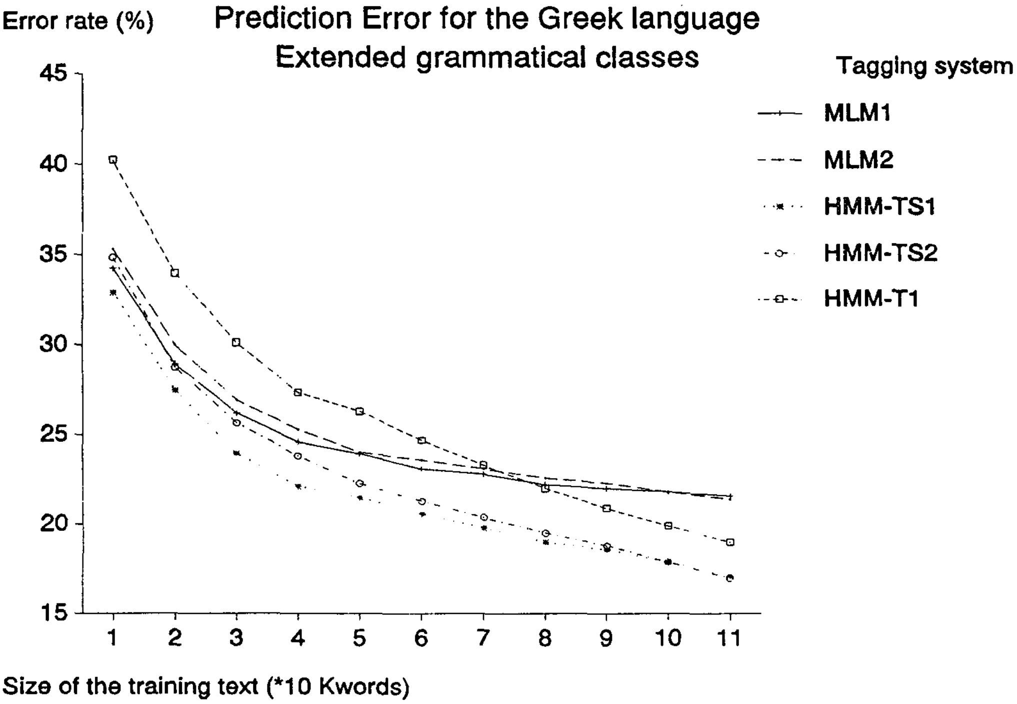


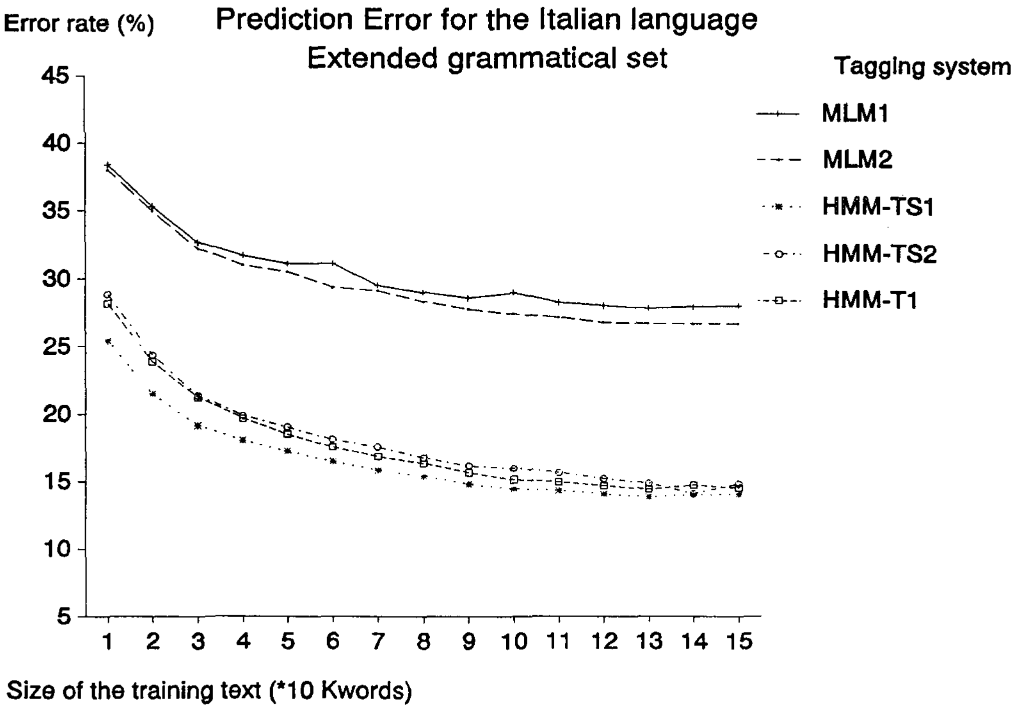


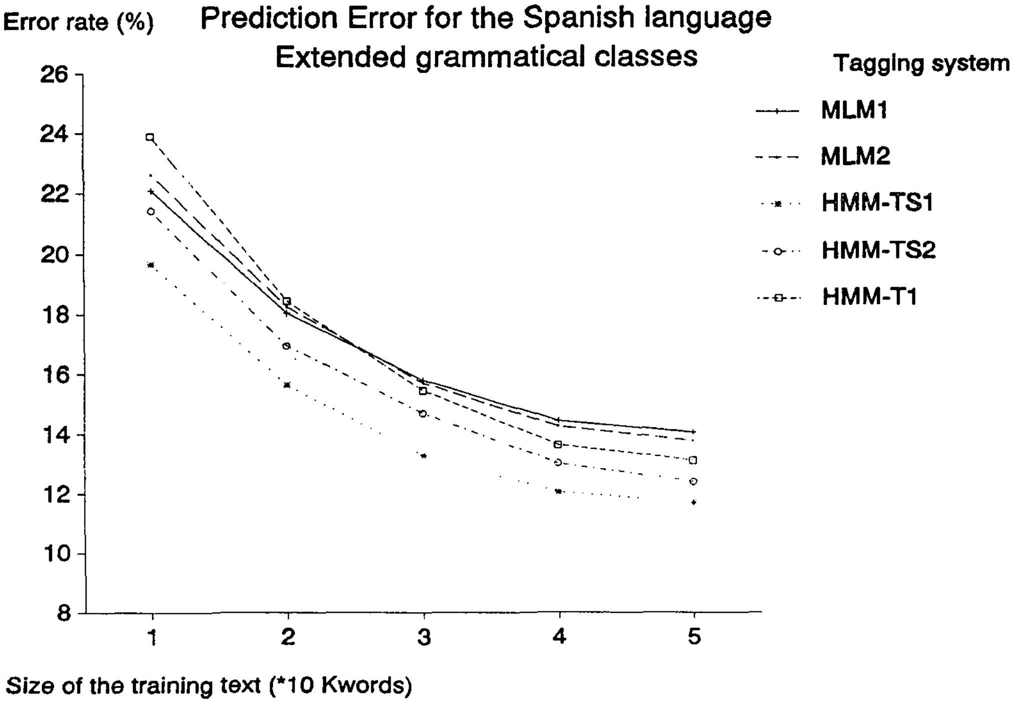

Error rate (%)
Prediction Error for the English EEC-law text
Extended grammatical classes
Tagging system
MLM1
MLM2
HMM-TS1
HMM-TS2
HMM-T1
13
12
11
10
9
8
7
6
5
4
3
2
1
2
3
4
5
6
7
8
9
10
Size of the training text (*10 Kwords)


Error rate (%)
Prediction Error for the French language
Extended grammatical classes
Tagging system
MLM1
MLM2
HMM-TS1
HMM-TS2
HMM-T1
16
15
14
13
12
11
10
9
8
7
6
5
4
1
2
3
4
5
6
7
8
9
Size of the training text (*10 Kwords)


Error rate (%)
Prediction Error for the German language
Extended grammatical classes
28
26
24
22
20
18
16
14
12
10
8
1
2
3
4
5
6
7
8
9
Size of the training text (*10 Kwords)
Tagging system
MLM1
MLM2
HMM-TS1
HMM-TS2
HMM-T1


Error rate (%)
Prediction Error for the Greek language
Extended grammatical classes
45
40
35
30
25
20
15
1
2
3
4
5
6
7
8
9
10
11
Size of the training text (*10 Kwords)
Tagging system
MLM1
MLM2
HMM-TS1
HMM-TS2
HMM-T1


Error rate (%)
Prediction Error for the Italian language
Extended grammatical set
Tagging system
MLM1
MLM2
HMM-TS1
HMM-TS2
HMM-T1
45
40
35
30
25
20
15
10
5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Size of the training text (*10 Kwords)


Error rate (%)
Prediction Error for the Spanish language
Extended grammatical classes
Tagging system
MLM1
MLM2
HMM-TS1
HMM-TS2
HMM-T1
26
24
22
20
18
16
14
12
10
8
1 2 3 4 5
Size of the training text (*10 Kwords)

References

Brill, E. (1992). "A simple rule-based part of speech tagger." In *Proceedings, Third Conference on Applied Natural Language Processing*. Trento, Italy, 152–155.

Cerf-Danon, H., and El-Beze, M. (1991). "Three different probabilistic language models: Comparison and combination." In *Proceedings, International Conference on Acoustics Speech and Signal Processing*, 297–300.

Charniak, E.; Hendrickson, C.; Jacobson, N.; and Perkowitz, M. (1993). "Equations for part-of-speech tagging." In *Proceedings, National Conference on Artificial Intelligence*.

Church, K. (1988). "A stochastic parts program and noun phrase parser for unrestricted text." In *Proceedings, Second Conference on Applied Natural Language Processing*. Austin, Texas, 136–143.

Church, K., and Gale, W. (1991). "A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams." *Computer Speech and Language* 5, 19–24.

Cutting, D.; Kupiec, J.; Pederson, J.; and Sibun, P. (1992). "A practical part-of-speech tagger." In *Proceedings, Third Conference on Applied Natural Language Processing*. Trento, Italy, 133–140.

Dermatas, E., and Kokkinakis, G. (1988). "Semi automatic labelling of Greek texts." In *Proceedings, Seventh FASE Symposium SPEECH '88*. Edinburgh, 239–245.

Dermatas, E., and Kokkinakis, G. (1993). "A system for automatic text labelling." In *Proceedings, Eurospeech-90*. Paris, 382–385.

Dermatas, E., and Kokkinakis, G. (1993). "A fast multilingual probabilistic tagger." In *Proceedings, Eurospeech-93*. Berlin, 1323–1326 (presented also in the Eurospeech-93 exhibition).

Dermatas, E., and Kokkinakis, G. (1994). "A multilingual unlimited vocabulary stochastic tagger." In *Advanced Speech Applications—European Commission ESPRIT (1)*, edited by K. Varghese, S. Pfleger, and J. Lefevre, 98–106. Springer-Verlag.

Eineborg, M., and Gamback, B. (1993). "Back-propagation based lexical acquisition experiments." In *Proceedings, NeuroNimes: Neural Networks and their Industrial & Cognitive Applications*. Nimes, 169–178.

Elenius, K. (1990)."Comparing a connectionist and rule based model for assignment parts-of-speech." In *Proceedings, International Conference on Acoustics, Speech and Signal Processing*, 597–600.

Elenius, K., and Carlson, R. (1989). "Assigning parts-of-speech of words from their orthography using a connectionist model." In *Proceedings, European Conference on Speech Communication and Technology*. Paris, 534–537.

Partners of ESPRIT-291/860 (1986). "Unification of the word classes of the ESPRIT Project 860." BU-WKL-0376. Internal Report.

Essen, U., and Steinbiss, V. (1992). "Cooccurrence smoothing for statistical language modelling." In *Proceedings, International Conference on Acoustics, Speech and Signal Processing*, 161–164.

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

He, Y. (1988). "Extended Viterbi algorithm for second order hidden Markov process." In *Proceedings, International Conference on Acoustics, Speech and Signal Processing*, 718–720.

Jacobs, P., and Zernik, U. (1988). "Acquiring lexical knowledge from text: A case study." In *Proceedings, Seventh National Conference on Artificial Intelligence*. Saint Paul, Minnesota, 739–744.

Jardino, M., and Adda, G. (1993). "Automatic word classification using simulated annealing." In *Proceedings, International Conference on Acoustics, Speech and Signal Processing*, 41–44.

Karlsson, F. (1990). "Constraint grammar as a framework for parsing running text." In *Proceedings, Thirteenth International Conference on Computational Linguistics*. Helsinki, Finland, 168–173.

Karlsson, F.; Voutilainen, A.; Anttila, A.; and Heikkila, J. (1991). "Constraint grammar: A language-independent system for parsing unrestricted text, with an application to English." In *Workshop Notes from the Ninth National Conference on Artificial Intelligence*. Anaheim, California.

Katz, S. (1987). "Estimation of probabilities from sparse data for the language model component of a speech recognizer." *IEEE Trans. on Acoustics, Speech, and Language Processing*, 35(3), 400–401.

Kupiec, J. (1992). "Robust part-of-speech tagging using a Hidden Markov Model." *Computer Speech & Language*, 6(3), 225–242.

Maltese, G., and Mancini, F. (1991). "A technique to automatically assign parts-of-speech to words taking into account word-ending information through a probabilistic model." In *Proceedings, Eurospeech-91*, 753–756.

Marcus, M.; Santorini, B.; and Marcinkiewicz, M. (1993). "Building a large annotated corpus of English: The Penn Treebank." *Computational Linguistics*, 19(2), 315–330.

McInnes, F. (1992). "An enhanced interpolation technique for context-specific probability estimation in speech and language modelling." In *Proceedings, International Conference on Spoken Language Processing*, 1491–1494.

Merialdo, B. (1991). "Tagging text with a probabilistic model." In *International Conference on Acoustics, Speech and Signal Processing*, 809–812.

Merialdo, B. (1994). "Tagging English text with a probabilistic model." *Computational Linguistics*, 20(2), 155–171.

Meteer, M.; Schwartz, R.; and Weischedel, R. (1991). "Empirical studies in part of speech labelling." In *Proceedings, Fourth DARPA Workshop on Speech and Natural Language*. Morgan Kaufman.

Nakamura, M.; Maruyama, K.; Kawabata, T.; and Shikano, K. (1990). "Neural network approach to word category prediction for English texts." In *Proceedings, Thirteenth International Conference on Computational Linguistics*. Helinski, Finland, 213–218.

Pelillo, W.; Moro, F.; and Refice, M. (1992). "Probabilistic prediction of parts-of-speech from spelling using decision trees." In *Proceedings, International Conference on Spoken Language Processing*, 1343–1346.

Rabiner, L. (1989). "A tutorial on Hidden Markov Models and selected applications in speech recognition." In *Proceedings, IEEE* 77(2), 257–285.

Tao, C. (1992). "A generalisation of discrete Hidden Markov Model and of Viterbi algorithm." *Pattern Recognition*, 25(11), 1381–1397.

Voutilainen, A., and Tapanainen, P. (1993). "Ambiguity resolution in a reductionistic parser." In *Proceedings, Sixth Conference of the European Chapter of the Association for Computational Linguistics*. Utrecht, Netherlands, 394–403.

Voutilainen, A.; Heikkila, J.; and Antitila, A. (1992). "Constraint grammar of English." Publication 21, Department of General Linguistics, University of Helinski, Helinski, Finland.

Watson, B., and Chung Tsoi, A. (1992). "Second order Hidden Markov Models for speech recognition." In *Proceedings, Fourth Australian International Conference on Speech Science and Technology*, 146–151.

Weischedel, R.; Meteer, M.; Schwartz, R.; Ramshaw, L.; and Palmucci, J. (1993). "Coping with ambiguity and unknown words through probabilistic models." *Computational Linguistics*, 19(2), 359–382.

Wothke, K.; Weck-Ulm, I.; Heinecke, J.; Mertineit, O.; and Pachunke, T. (1993). "Statistically based automatic tagging of German text corpora with parts-of-speech—some experiments." TR75.93.02-IBM. IBM Germany, Heidelberg Scientific Center.