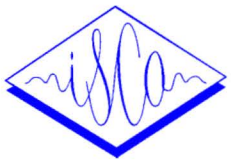

First European Conference on
Speech Communication and Technology
EUROSPEECH '89
Paris, France, September 27-29, 1989

WORD ENDINGS ANALYSIS OF EUROPEAN LANGUAGES

M.REFICE, M.SAVINO

Istituto di Scienze dell'Informazione - Universita' di Bari
Via Amendola, 173 - 70126 Bari - Italy

ABSTRACT

This paper reports the preliminary results of a study conducted upon six European languages and started within the framework of the Esprit Project 860 "Linguistic Analysis of the European Languages".
A software tool allows to dynamically define the length of endings according to some specified constraint with the purpose of identifying POS cohorts.
Some of the statistical results obtained from extensive runs performed upon the given corpora are reported and briefly discussed.
The performances of the identified POS cohorts with the respect to a labelling task, have been also assessed and the main outcomes are presented and discussed.

INTRODUCTION

In a multilingual text-to-speech and speech-to-text conversion system, based on large vocabularies, a feasible approach may consist on trying to implement simple and effective methods which are as common as it is possible for all the considered langauges [1].

Accordingly, statistically derived information may prove to be as effective as specific linguistic knowledge.

The main purpose of this study was to derive statistical properties, namely Part Of Speech (POS) cohorts, from the analysis of word endings for six european languages.

In such a system the assignement of a POS label to a word may be crucial as replacement of the dictionary for words not belonging to it, as well as an independent knowledge source able to help in any grammatical disambiguation task.

LEARNING CORPORA

We have been analysing the dictionaries which have been automatically derived from labelled texts consisting of about 100.000 words per each of the languages; each dictionary contains about 10.000 words.

The texts consist of an issue of the CEC official journal and therefore refer to the same domain for all the languages.

This allows a reasonable comparison among the considered languages, namely Dutch, English, French, German, Italian and Spanish, eventhough each text cannot be considered as representative of its corresponding language.

Although the dictionaries contain more information, namely number of occurrences and phonetic transcription for each word, this analysis has only taken into account the graphemic word and its grammatical category.

The category set which has been assigned to the words belonging to the dictionaries is a very detailed one and comes from the unification of word classes for all the six mentioned languages [2].

A specially designed software module allows the "mapping" of the detailed classes into any set of "hyperclasses" called "cover-symbols" at the user's choice.

ANALYSIS

An ending is dynamically defined as a string of characters with a constraint of a maximum of 8 characters or half minus one of the word length [3].

A POS cohort consists then of all the words sharing the same ending and belonging to the same POS; since for each ending several cohorts may exist, a rule is defined when the population of a cohort prevails upon the others of more than a given threshold.

In the following of this report,POS refers to the Main grammatical category, namely Verb, Noun, Adjective, Adverb, Pronoun, Preposition, Article/Determiner, Conjuction, Particle, Interjection, Miscellaneous [2], which have been obtained by mapping the detailed grammatical categories contained in the dictionaries.

In all the examined languages, for values of the threshold higher than 30%, the number of obtained rules is linearly dependent on the value of the threshold; the coefficient is about the same for all the languages.

The numbers of rules for two values of the threshold are shown in Fig.1 for all the languages.

LANGUAGE	THRESHOLD 20%	THRESHOLD 95%
DUTCH	142	694
ENGLISH	94	549
FRENCH	179	671
GERMAN	47	949
ITALIAN	24	533
SPANISH	32	916

Fig. 1 Number of rules for two values of the threshold.

The linear dependence of the number of rules on the threshold stems from the fact that, whereas the same rule does exist for more than one ending, only the rule which refers to the shortest ending is saved.

As a consequence of this choice, endings may vary with the adopted threshold and only some of them remain in the same cohort as the rules costraint changes.

The weighted mean length of the obtained endings for three values of the threshold is shown in Fig. 2, along with the weighted mean length of the words belonging to the dictionaries.

RULES TESTING

In order to assess the performance of the rules obtained from the word endings analysis, we have been testing the rules in a labelling task in self-consistency mode.

As it may be expected, the higher the threshold and therefore the more stringent is the constraint of a rule definition, the more correct is the corresponding labelling but the less is the chance for a rule to be applied in any given text.

LANGUAGE	THRESHOLD 20%	THRESHOLD 50%	THRESHOLD 80%	WORDS MEAN LENG.
DUTCH	3.4	3.7	4.4	9.9
ENGLISH	2.6	2.9	3.5	7.8
FRENCH	3.1	3.5	4.1	8.6
GERMAN	1.9	3.9	4.4	10.8
ITALIAN	1.0	3.3	4.1	8.6
SPANISH	1.3	3.4	4.2	8.8

Fig. 2 Mean length of endings and words.

We have therefore defined the following two parameters, able to describe the phenomenon:

- Correctness, as the number of right labels assigned by the set of rules with respect to the total occurrences of rule applications;

- Coverage, as the total number of rule application occurrences with respect to the total number of words in the text.

Fig. 3 reports the results of the mentioned labelling task for the six languages.

The first inspection of the graphs shown in Fig.3 indicates two main trends: Dutch, English, French, Italian and Spanish behave almost the same with respect to the mentioned parameters while German shows a quite different situation.

In the first case, Correctness and Coverage are both correlated with the threshold but in opposition one to the other; depending on the specific application one may look for, it may be defined an optimum value of the threshold as the point where the two lines cross.

For German, Coverage is almost independent of the threshold and no evidence seems to be for its possible increase.

As it has been shown by other studies [4], in a lexical access task forward prediction seems to be more successful than backward prediction when Germanic languages are taken into account.

For we have been deriving rules for German word beginnings using the same learning corpus; the result of the corresponding test is shown in Fi. 4 and its behaviour seems to be very similar with respect to the endings for the other languages.

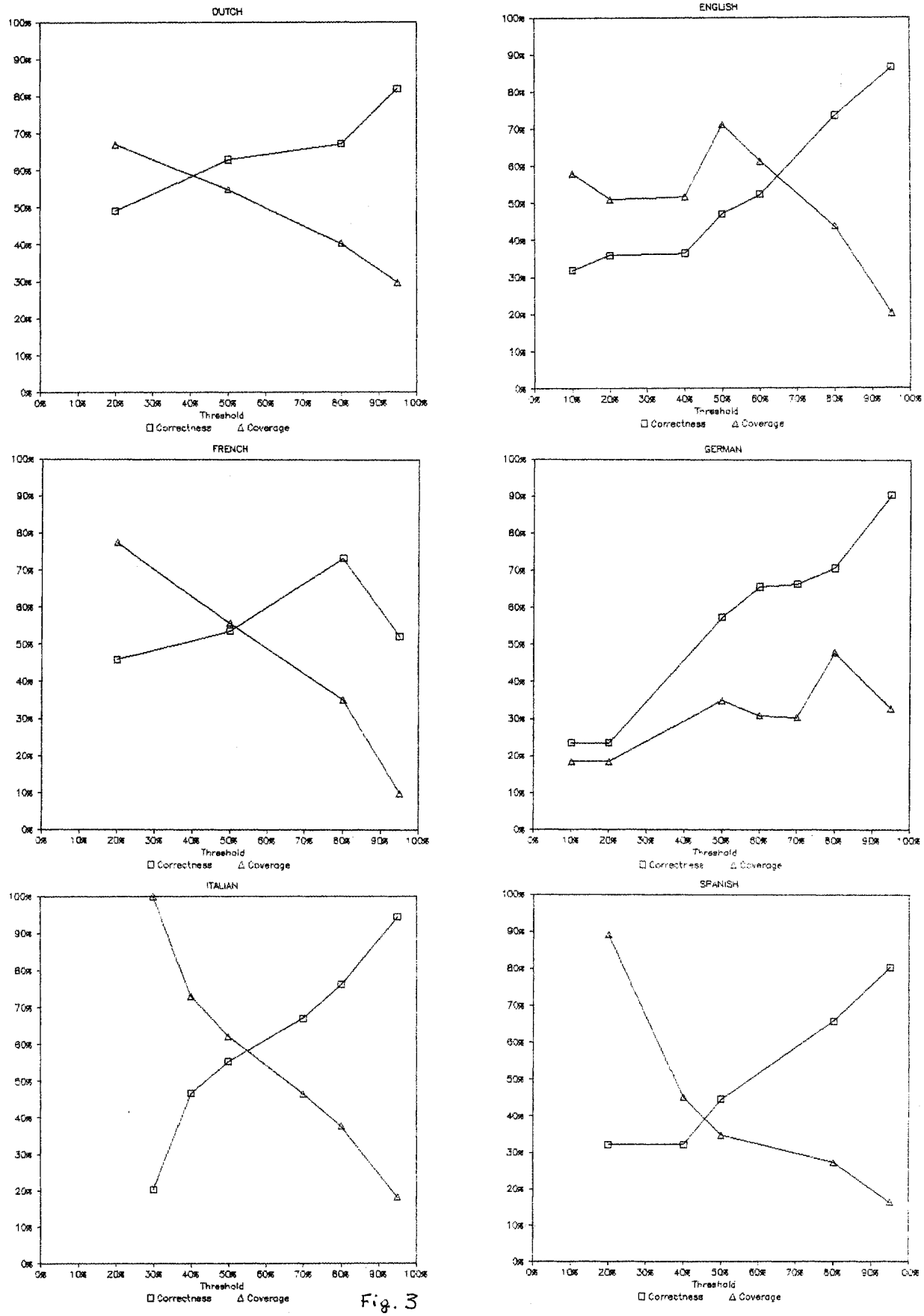
DUTCH
ENGLISH
FRENCH
GERMAN
ITALIAN
SPANISH
Threshold
Correctness
Coverage

Fig. 3

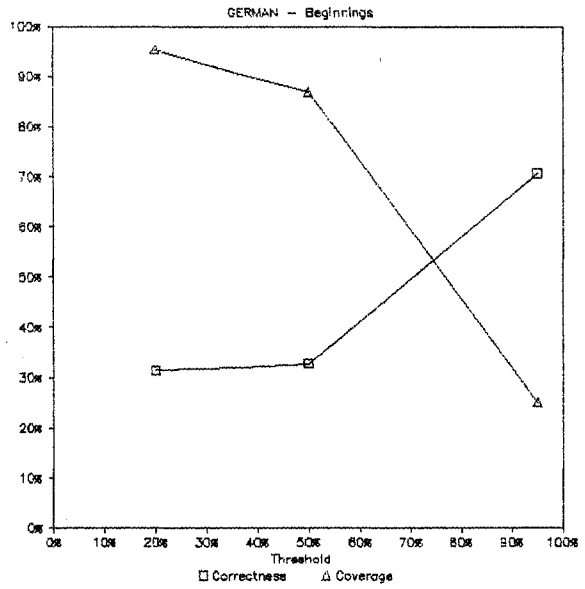

Fig. 4 Correctness and Coverage for German beginnings.

REFERENCES

[1] L.Boves, M.Refice, "The linguistic processor in a multi-lingual text-to-speech and speech-to-text conversion system", Proceedings of the European Conference on Speech Technology, Eds. J.Laver and M.A.Jack, CEP Consultans Ltd, Edinburgh, 1987

[2] M.Kugler-Kruse, "Unification of word classes of the Esprit project 860",Last revision 22/2/1989, Esprit 860 internal report BU-WKL-0376

[3] A.Mastrolonardo, M.Savino, "Word endings analysis", Esprit 860 internal report CS-RP0131, 31/3/1989

[4] R.Carlson, K.Elenius, B.Granstrom, S.Hunnicut, "Phonetic and ortographic properties of the basic vocabulary of five European languages" French-Swedish Seminar, Grenoble, April 1985