# Mining Subcategorization Information by Using Multiple Feature Loglinear Models

**Nuno C. Marques**[1,2] and **Gabriel Pereira Lopes**[2] and **Carlos A. Coelho**[3]

nmm@di.fct.unl.pt and gpl@di.fct.unl.pt and coelho@isa.utl.pt

[1] DCET - Universidade Aberta
[2]CENTRIA- Dept. Informática - FCT/UNL
[3]Dept. Matemática - ISA/UTL

## Abstract

In this paper we show how several non-independent features can be conjugated for loglinear statistical modeling of subcategorization information. Having this in mind we will present a method for unsupervised learning of statistical loglinear models for words with the same subcategorization frame, using huge collections of fully automatically part-of-speech tagged texts. Experiments done show that in some cases we can improve our previous approach precision (using just one feature) from 82% to 91%.

## 1 Introduction

In this paper we show how several non-independent features can be conjugated in order to quantitatively characterize word subcategorization frames. Elsewhere, (Marques et. al. 1998a), (Marques et. al. 1998b) we have showed that loglinear modelling (Agresti 1990) can be used for clustering verbs (and other words), based on the occurrence of a single relevant feature extracted from corpora. However, there were cases, such as subject verb (SV) order inversion, VS, that led, in Portuguese, to incorrect classification of some intransitive verbs as transitive ones, taken as subcategorizing a noun phrase. A fully automatically POS-tagged corpus was then used as an entry for a system capable of clustering (in an unsupervised manner) the words requiring as their arguments (subcategorizing) the same kind of syntactical structures. Here we will show how several interacting features (statistically non-independent) may be conjoined in a single loglinear statistical model.

Several authors have already presented work on subcategorization extraction in English. Michael Brent (Brent 1993) presented an approach where each subcategorization frame could be introduced by a small set of morpho-syntactic cues. For instance, the presence of the pronoun *me* is a good cue for the presence of an object. Brent's method was based on trying to find verbs co-occurring with such cues frequently enough for making evident the presence of a subcategorization class. However, Brent's cues were very restrict and specific: only highly accurate, but extremely rare cues were used. Manning (Manning 1993) and, more recently, Ted Briscoe et alia (Briscoe Carroll 1997), made a much better use of available resources. By using a part-of-speech tagger and a grammar (a simple finite state grammar in Manning's case and a wide coverage partial parser in Briscoe and Carroll's case) they used complex POS-tags (noun phrases, prepositional phrases, etc.) to replace Brent's cues. However just the type of cues used were changed by both these authors. The statistical method used to select relevant cues remained the same.

In (Marques et. al. 1998a), (Marques et. al. 1998b), loglinear modeling allowed us to maximize the available resources, classifying all verbs occurring more than 80 times in a corpus, for most of the subcategorization classes, a value that Brent's methodology couldn't keep up to. By using a smaller corpus, we were able to extract subcategorization information for more verbs. Moreover, while Brent's method was only able to assign a binary information regarding verbal subcategorization, loglinear modeling gives us quantitative information regarding the verbal subcategorization. Until now, (Marques et. al. 1998a), (Marques et. al. 1998b) we have presented loglinear models capable of using just one cue. However, we found that, in some cases, it can be advisable to model several cues with the same loglinear model. In this article we are going to present situations where the use of a

single cue is insufficient. These examples will allow us to show how the basic loglinear model can be extended in order to handle several features simultaneously. In order to do so, we will focus on the problem of verbal transitivity, since it was the one that required the use of more features, in order to describe rather specific linguistic phenomena: the main problem was, as we mentioned earlier, the inversion of the traditional Subject-Verb order that characterizes some Portuguese verbs.

After a brief description of the framework we developed for learning subcategorization by using loglinear models, we will comparatively study three possible feature sets. Several loglinear models with and without interactions among features and scores will be used. The results of the clustering process will be evaluated by comparing the results automatically obtained with an independent classification presented in commercial dictionaries. Improvements were particularly noticeable in intransitive verbs where the automatic classification precision was raised from 82% to 91%. Conclusions will be drawn regarding the advantages and problems of introducing more features in a particular loglinear model. The acquired results are presently paving the way to new parsing methodologies, capable of automatically overcome either the lack of subcategorization information or a default subcategorization frame assignment to every kind of word in the lexicon.

## 2 LogLinear Models for Mining Subcategorization Classes

In table 1, in the second column we present the frequencies of the article-noun pair (i.e., the total number of verbal forms immediately followed by an article, followed by a noun) for verbs *afirmar* (to assert) and *encontrar* (to meet), in the first column. In the third column, it is presented the frequency of verbal forms followed by an article, which is not followed by a noun. The fourth column shows the remaining occurrences of the verb which are not followed by article.

This table is called a contingency table. Columns represent feature counts and rows the verbs chosen for analysis. The statistical relations between the rows and columns in such a table can be analysed by using loglinear models (Agresti 1990). In (Marques 2000) it was shown

| | $(art,n)$ | $(art,\overline{n})$ | $\overline{art}$ | $\widehat{\lambda^X}$ |
|---|---|---|---|---|
| $v_{afirmar}$ | 514 | 379 | 7290 | 0 |
| $v_{encontrar}$ | 413 | 320 | 6092 | $-0.1815$ |
| $\widehat{\lambda^Y}$ | 0 | $-0.2823$ | 2.670 | $\widehat{\lambda}$=6.225 |

Table 1: A contingency table for two verbs and three attributes.

that if a set of verbs (rows) in a contingency table have an independent behaviour regarding a chosen set of features, then those verbs should present the same subcategorization class. This approach has the advantage of taking into account several verbs simultaneously. That allows us to determine a subcategorization class using both less accurate cues and less frequent verbs than what could be done by using the systems presented by previous authors (Brent 1993),(Manning 1993),(Briscoe Carroll 1997). By clustering verbs, we were able to build a system where almost no grammar knowledge is assumed (we simply use the part-of-speech tags automatically assigned by a POS-Tagger (Marques 2000)), and, where the available text is very efficiently used. As a result, we are able to propose a new quantitative and statistically well-founded view of the subcategorization process. In cases where the independence model is applicable, the expected value for the observed counts in a contingency table could be estimated using the independence loglinear model (Agresti 1990):

$$(1) \qquad logE_{ij}=\lambda+\lambda_i^X+\lambda_j^Y \qquad (i=1,...,I;j=1,...,J),$$

$logE_{ij}$ is the logarithm of the expected frequency of cell $(i,j)$ and equals the sum of a constant $\lambda$ with a row parameter $\lambda_i^X$ and a column parameter $\lambda_j^Y$. The estimated values of these parameters are represented respectively in the right column (headed by $\widehat{\lambda^X}$) and lower row (headed by $\widehat{\lambda^Y}$). In models like this one, direct estimators can be inferred from data. In more complex cases a statistical package (Healy 1988) is used to fit the loglinear independence model to our data (Marques 2000), (Agresti 1990).

We can also evaluate how good a model fits the available data by comparing the estimated values with the real ones. We will use the likelihood-ratio statistic:

$$(2) \qquad G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} O_{ij}\log(\frac{O_{ij}}{E_{ij}}),$$

where $O_{ij}$ is the observed frequency for cell $(i, j)$. When a model holds, this statistic has a large-sample chi-squared distribution with $(I-1)(J-1)$ degrees of freedom. In the above example, $G^2 = 0.357322$ is a value well below 5.991476 (the $95^{th}$ quartile of the chi-squared distribution with two degrees of freedom), i.e. we could not reject the independence assumption and so, the two verbs should be considered as belonging to the same subcategorization class.

In (Marques et. al. 1998b), we have first shown how loglinear models, evaluated by $G^2$, can be used to find independent verb clusters: if we have a set of features $F_1, F_2, ..., F_r$, a cluster of verbs $\overrightarrow{v_1}$ and a candidate verb $v_2$, by modeling the contingency table $X =< F_1, F_2, ..., F_r >$, $Y =< \overrightarrow{v_1}, v_2 >$, we will be able to decide if verb $v_2$ has the same behavior with respect to all of the features $F_1, F_2, ..., F_r$ as the group of verbs $\overrightarrow{v_1}$.

After applying this algorithm to a set of verbs extracted from a previously automatically POS-tagged corpus three independent, commercial dictionaries, using the more traditional notion of subcategorization ((Buse 1994), (Ventura Caseiro 1992) and (Silva Tavares 1989)), were used for evaluating how much did our model approached the traditional notion of verbal subcategorization. Table 2 presents the results using the presence or absence of the article directly after the verb as a single feature in a independence model (labeled $[ART]$ for further reference) in terms of precision and recall[1].This experiment will work as a quality control. In this table, the first two columns represent values in the corpus, that is, how many errors are found in a corpus when using the given classification (this measure is relevant for parsers). The second pair of columns gives another perspective for the same problem — how many verbs disagree with the cluster type. We have tagged each cluster into transitive (*elem* column) or intransitive (*dont* column) [2], according to the subcategorization assigned by the dictionary to the first verb in the cluster (the most frequent one). Based on this cluster classification we compute precision (how many verbs, in the corpus and in the dictionary, disagree with the cluster heading verb) and recall values (how many verbs with a given classification disagree with the cluster heading verb), as a value pair. The lower value shows a pessimistic measure that does not take into account the verb used to classify the cluster. Yet since we are using an unsupervised method, this evaluation seemed to be too restrictive, so we have also included an optimistic measure: the value that takes into account the verb used to classify the cluster. The real value should be between these two. In the following experiments, only verbs with a total number of occurrences greater or equal to 81, and, also, the pairs of attributes with more than 3 occurrences were considered. In the last line it is represented how many clusters were obtained for the 1279 verbs classified: 9 clusters for intransitive verbs; 53 clusters for verbs subcategorizing a noun phrase; and 12 clusters headed by verbs that do not appear in the commercial dictionaries.

## 3 Problems of Using Just One Feature

During the study of the use of loglinear models for mining subcategorization classes (Marques 2000), transitive verbs showed us a particularly important problem, not covered by the previous experiment: the inversion of the subject-verb order. This inversion occurs due to a strong trend, in Portuguese language, to place the subject after the verb, particularly in the case of intransitive verbs. Indeed, while in the majority of the previous cases only one feature (and its complement, presence or absence of that feature) would be enough to characterize all the occurrences of a syntactic argument within a given subcategorization class, in the case of transitive verbs this is not necessarily the case. Moreover, while, as in the previous experiment, the

[1] Although results are different from the ones presented in (Marques et. al. 1998b) (due to a more rigorous selection and classification of the evaluation dictionary, please see (Marques 2000)), the conclusions and overall behavior are equivalent to the ones in (Marques et. al. 1998b).

[2] More generally, we use *elem* in order to mark the presence of any given subcategorization and *dont* to mark its absence.

| | corpus | | dictionary | |
|---|---|---|---|---|
| | dont | elem | dont | elem |
| Precision | 82%-62% | 99%-97% | 59%-52±13% | 98%-97±2% |
| Recall | 93%-81% | 96%-92% | 80%-75±13% | 94%-93±2% |
| total | 319749 | 1794951 | 49 | 463 |
| N | 9+53(+12)/1279 | | | |

Table 2: Results of the clustering with the attribute ARTICLE.

presence of the article can be used as being the only feature for presence of a particular phrase (namely a noun phrase), some other times, the absence of article occurs:

- a Escola Prática de Infantaria vai **realizar exercícios** de fogo real durante a próxima semana (the Practical School of Infantry will **practice** 1:[real fire] **exercises** [1] next week).

In this example, the verb *realizar (to practice, to do, ...)* occurs immediately followed by a noun ( *exercícios (exercises)*). For instance, in our *corpus*, the verb *to carry* is 1741 times followed by an article, and 660 times followed by a noun. In extreme cases the verb can even have a higher frequency of the attribute noun after the verb than the frequency of the article after the verb (this is the case of verbs as *autorizar (to authorize, ...)* or *fazer (to make))*. In this case, either the noun presence, or the article presence (both of them immediately after the verb), function as a valid cue for the presence of a noun phrase argument after the verb.

During our analysis of the errors made by the clustering algorithm using only the *article* cue, we have also found numerous cases of inversion of the subject order, namely with verb *ocorrer (to occur)*:

- Sexta-feira *ocorreu* o massacre (...) (*Friday, 1: [the slaughter] occurred [1] (. . . )*)
- *Ocorreu* um incidente violento entre os agentes da PJ (...) (*1:[a violent incident among the PJ agents] occurred [1] (...)*).

However, in these cases there is a restriction that can be used to detect this inversion: the agreement in number between the verb and the word that follows it (here denoted by *agr*). As a matter of fact, subject and verb must agree in number, independently of the subject-verb order. This is not the case of an object. In the case of the row effects model (see next section), for example, the verb *morrer (to die)* is used in the definition of a cluster containing transitive verbs *obter (to obtain, to get, ...)*, *congregar (to congregate, ...)* and *enfatizar (to emphasize, ...)*. Thus, when a frequent verb is followed by its subject, it can be expected a larger count of number agreement between the verb and its subject and than in the case of a normal object following its verb. So, number agreement will therefore be, a third attribute we want to study in this paper.

We will use these three cues (*art*, *n* and *agr*) to illustrate how the loglinear models can be used successfully by conjoining several attributes. As in the previous case, we will follow an attribute by its complement. In the case of attributes used for denoting the presence of the article after the verb we have represented the counts by *art* and $\overline{art}$. In the next experiment we will try to conjugate this attribute, with the attribute presence of noun after the verb ($n$). Thus a categorical variable with 3 attributes will be used: the first one will be the presence of article after the verb ($art$), the second, the presence of a noun after verb ($n$) and finally, as in previous case, the last attribute will count all the remaining cases (in the case, it will be represented by $\overline{art \vee n}$). We will analyse some possible models for the study of this case, from the model of simple independence to a row effects model with scores(Agresti 1990).

Finally we will study the agreement between the verb and the word that follows it (represented by variable $AGR$) as a distinct categorical variable for the study of transitive verbs.

Table 3, presents a baseline for the accuracy for each one of the selected attributes.

| | Art | $Art \vee N$ | $Art - Agr$ | $(Art \vee N) - Agr$ |
|---|---|---|---|---|
| Total of verbs with cue frequency> 80 | 1279 | 1086 | 555 | 336 |
| Total of verbs with cue frequency $< 4$ | 85 | 278 | 809 | 1028 |
| More frequent class | elem | elem | elem | elem |
| Baseline precision in corpus | 83% | 83% | 85% | 87% |
| Baseline precision in the dictionary | $87 \pm 3\%$ | 88±3% | 88±3% | 90±3% |

Table 3: Baseline precision (using just dictionary lookup) of the text, for the studied attributes. Second row tells how many verbs where excluded due to small counts ($< 4$).

## 4 Results with the attributes Article and Noun

In this experiment we will join attributes $n$ and $art$, representing the presence of a noun or an article immediately after the verb. This attributes can be represented by a statistical variable ($ART \vee N$). Thus, the two attributes had been joined in a single categorical variable. For example, in the case of the verbs *considerar (to consider, ...)* and *votar (to vote)* we can construct a contingency table, such as the one presented in the table 4.

The use of more than two attributes in a categorical variable enables the use of the so called row effects model (Agresti 1990). In this experiment the results of the process of clustering of data will be compared using the simple model of independence (see table 2) with the results obtained for a row effects model (Agresti 1990). Table 6 presents the results obtained by using the row effects model. In this model we have inserted a relative score for each attribute, in order to represent a relative ordering of these attributes. The attributes article and noun both being pointers for presence of noun phrases, we have assigned a score of 1 to the attribute article ($art$), and a score of 2 to the attribute noun ($n$)[3]. The complement column received score 15 (a value that we experimentally found to be distant enough from the other two attributes, that also shows how we can insert subjective judgements about feature importance in the model). Since there are no direct estimators for the parameters of the row effects model, we used the statistical package *Glim∗* (Healy 1988) to make the iterative adjustment of the parameters $\lambda$, $\lambda^V$, $\lambda^{ART \vee N}$ and $u_{art \vee N}\lambda^V$ of the model

[3]Since assigning them equal scores (assigning, for example, score 1 to both the attributes $art$ and $n$), would be equivalent to the use of the independence model with only 2 attributes.

| | $art_1$ | $n_1$ | $\overline{art_1 \vee n_1}$ |
|---|---|---|---|
| to consider | 3279 | 413 | 18906 |
| to vote | 584 | 74 | 3227 |

Table 4: Contingency table conjugating the attributes article or noun (with its complement) immediately after the verb

$$\log E_{v,j} = \lambda + \lambda_v^V + \lambda_j^{ART \vee N} + u_j \lambda_v^V, \quad (3)$$

with $u_{art} = 1, u_n = 2, u_{\overline{art \vee n}} = 15$, in accordance with the observed data.

If we compare the results of the independence model for the cues $art$ and $n$ (table 5) with the results of the independence model using only the cue $art$, we immediately notice an increase in the number of clusters needed to describe the available information. Thus, a total of 74 (9 + 53 + 12) clusters in the description of 1279 verbs (approximately 17 verbs per cluster) was increased to 212 clusters (20+133+59) for the description of 1086 verbs (approximately 5.12 verbs per cluster). Although the results in terms of accuracy and recall have also improved, in this case, the number of clusters manifestly seems to be too high. This result was expected, since it was required that the model would describe one more attribute, without increasing model descriptive power . When applying the row effects model with column scores, we have increased the descriptive capacity of the model. In this case we can notice a strong reduction in the number of clusters needed to describe our

| | corpus | | dictionary | |
|---|---|---|---|---|
| | dont | elem | dont | elem |
| Precision | 88%-60% | 99%-96% | 73%-60±15% | 98%-97±2% |
| Recall | 95%-80% | 98%-91% | 84%-74±2% | 96%-95±14% |
| total | 334366 | 1395246 | 55 | 451 |
| N | 20+133(+59)/1086 | | | |

Table 5: Results of clustering of the attributes article or noun after the verb, using the independence model.

| | corpus | | dictionary | |
|---|---|---|---|---|
| | dont | elem | dont | elem |
| Precision | 77%-52% | 99,5%-99% | 51%-43±13% | 99.8%-99.7±0,5% |
| Recall | 97%-91% | 95%-91% | 97%-96±8% | 94%-93±3% |
| total | 363924 | 1826947 | 33 | 473 |
| N | 9+52(+12) /1086 | | | |

Table 6: Results of the clustering of the attributes article or noun after the verb, using the row effects model with scores.

data: 73 $(9 + 52 + 12)$ clusters (14.9 verbs per cluster). Notice that, due to use of the row effects model, the number of clusters becomes smaller that the one acquired using the independence model.

The results seem slightly favorable after inserting the $n$ cue, when using the row effects model: the number of verbs classified as intransitive, that were incompatible with the dictionary was approximately equal (comparing the dictionary precision $43 \pm 13\%$ of correct verbs against $52 \pm 13\%$ when using the independence model $[ART]$) and the verbs classified as transitive verbs, improved ($99.7 \pm 0.5\%$ of accuracy after the introduction of attribute $n$ against only $97\% \pm 2$ when using only the attribute $art$).

Yet (as we would expect) the analysis of the errors made by the algorithm showed us that the problem of the inversion of the subject maintains. In the next experiment we will use the agreement cue to try to handle this problem.

## 5 Agreement in Number between the Verb and the Word to its Right

In most cases we can detect the agreement in number in Portuguese, by observing the ending of the verb. In our case we have used the sequences - am, -ão, - em, -eis or -mos for detecting a plural verb ending and ending -s for the plural article. We should stress that this rule is not a generic one, nor does it guarantee that all the agreement cases are caught. Thus with it, we can only detect some agreements in number between the verb and the eventual subject that follows the verb.

### 5.1 Addition of New Dimensions for the Study of Verbal Transitivity

In this experience instead of clustering the diverse attributes in only one variable, we have used one variable for each pair of attributes under study. Thus it will be necessary to construct contingency tables that represent all the combinations of attributes article or noun after the verb ($art$, $n$ and $\overline{art \vee n}$) and also the agreement between the word that follows the verb and the verb (i.e. the variable $AGR$, with the attributes $agr$ and $\overline{agr}$): ($art.agr$, $art.\overline{agr}$, $n.agr$, $n.\overline{agr}$, $\overline{art \vee n}.agr$ and $\overline{art \vee n}.\overline{agr}$). We will consider the variables $ART \vee N$ and $AGR$ dependent between themselves. We will also assume that the variable $ART \vee N$ is an ordinal variable. We will use the denomination $[ART \vee N \bowtie AGR]$ to represent the row effects model with dependencies between the variable $ART \vee N$ and $AGR$:

(4)
$$logE_{v,j,k} = \lambda + \lambda_v^V + \lambda_j^{ART \vee N} + \lambda_k^{AGR} + \lambda_j^{ART \vee N} \lambda_k^{AGR} + u_j \lambda_v^V \lambda_k^{AGR},$$

with $u_{art} = 1, u_n = 2$ and $u_{\overline{art \vee n}} = 15$. Once again, due to the inexistence of direct estimators for the parameters of the model, we used the Glim* package (Healy 1988) to do the iter-

ative adjustment of the model parameters. As an additional term of comparison we will also use the model $[ART, AGR]$ (that is the independence model between the cues *article* and *agreement*).

Some care has been taken in the study of less frequent verbs. Thus, the verbs selected for grouping with model $[ART]$, with no occurrences of agreement, have maintained the classification given by the model $[ART \vee N]$. The results of the clustering with model $[ART \vee N \bowtie AGR]$ are presented in the table 8 and the results with the model $[ART, AGR]$ are presented in the table 7. Once again, last line in tables denote the number of clusters obtained for verbs not subcategorizing a noun phrase (14 in table 8), the number of clusters headed by verbs subcategorizing a noun phrase (84 in table 8), and the number of clusters headed by verbs that do not occur in the commercial dictionaries we used for validating the results obtained (5, in the same table).

## 6 Analysis of the Results

We have noticed that the introduction of more attributes leads to an increase in the number of clusters needed to describe all the verbs. Thus, in the models with a lot of attributes, many of the studied verbs made up its own cluster with no other verbs in it. This analysis is initiated taking as example the verbs *concorrer (to concur, to submit, ...)*, *depender (to depend)*, *faltar (to lack, to miss, ...)*, *ocorrer (to ocurr, to happen, ...)*, *morrer (to die)*, *nascer (to born)* and *faltar (to lack, to miss, ...)* that were incorrectly clustered due to the inversion of the subject when clustered using only the attribute *art* (in the model $[ART]$). After inclusion of the cue agreement (in the model $[ART,AGR]$), only *nascer*(*to_born*) continued to be clustered with transitive verbs. Due to a joint frequency of the attributes $to\_born, art, agr$ of only 5 occurrences, this verb was clustered with *conseguir*(*to_achieve*), a verb that presents a high frequency (8505 occurrences against 3635 for *to_born*) and similar levels of agreement verb-argument:

- o Uzbequitão e o Japão, que conseguiram as suas primeiras medelhas de ouro (...). (*Uzbekistan and Japan, that had achieved their first gold medals*).

There was five verbs that, thanks to the introduction of the new attribute, have been used in the definition of new clusters. Thus *to_depend* was clustered with *to_answer*, and the verbs $to\_fit$ and *to_elapse* have been clustered with *to_grow*. The verb *to_occur* (and $to\_be\_a\_refugee$) were clustered with *to_disappear*, *to_concur* with *to_circulate* and, finally, *to_lack* was clustered with *to_declare_insolvent*. All these clusters were classified as verbs that do not subcategorize a noun phrase as their first argument. Among the referred verbs, *to_die* was not clustered with any other verb. So for all the verbs we are analysing, only *to_born* remained in a intransitive cluster. However the 212 clusters that were acquired are far from the ideal: although there was evidence in favor of the model, we still need improvements in order to try to get a bigger number of verbs per cluster.

As we could expect, the use of the attribute *noun*, in an independence model (table 5), confirms the comments already made: not being an attribute that contributes for the resolution of the problem of the inversion of the order of the subject, its only effect on the verbs under study is to diminish the size of the studied clusters. Thus, of the referred verbs, the only verbs that had correctly been clustered are the verbs *to_depend*, clustered with *to_belong* and $to\_fit$; and *to_born*, clustered only with the verb *to_correspond*. From remaining verbs, *to_die* is a defining verb that contradicts the other verbs that are clustered with it, while *to_elapse*, *to_lack*, *to_continue* and *to_occur* had incorrectly been grouped in clusters of transitive verbs. Once more, the huge number of clusters makes difficult to analyse the data.

In the row effects model with scores, due to the huge number of distinct patterns supported in the same cluster, the reverse effect was given: relatively to the respective independence models it had a remarkable reduction in the number of clusters. The analysis of the results became thus much easier. In practice, in the case of the model $[ART \vee N]$, none of the verbs that we were considering in this study (all of them intransitive verbs) was correctly clustered. All the verbs in study are enclosed in transitive clusters: *to_born* and *to_die* in the cluster headed by *to_carry_through*, *to_elapse*, *to_lack*

| | corpus | | dictionary | |
|---|---|---|---|---|
| | dont | elem | dont | elem |
| Precision | 97%-71% | 99.8%-99% | 81%-61±18% | 99%-99±1% |
| Recall | 99%-87% | 99.3%-97% | 92%-80±17% | 98%-96±2% |
| Total | 3392774 | 1713815 | 50 | 441 |
| N | 29+165(+17)/ 555 | | | |

Table 7: Clustering of the attributes article with agreement using the independence model.

| | corpus | | dictionary | |
|---|---|---|---|---|
| | dont | elem | dont | elem |
| Precision | 91%-61% | 99%-96% | 68%-41±23% | 98%-97±2% |
| Recall | 94%-68% | 99%-95% | 81%-58±28% | 96%-95±3% |
| Total | 2213490 | 1714433 | 26 | 283 |
| N | 14+84(+5)/ 336 | | | |

Table 8: Results of clustering the attributes article or noun and agreement using a row effects model with scores.

and *to_concur* with *to_continue*, *to_occurr* with *to_have_a_meting_with* and finally *to_depend* with *to_vote*. Indeed, the introduction of the score, only seems to have brought a slight improvement: several verbs with the same features (as *to_die* and *to_born*) are presented in the same cluster. This effect must probably be due to the incorporation of the attribute *noun*, that also has to have an inversion in the order SVO so that the model can adjust to the verbs in that cluster.

After the introduction of the cue agreement (and its complement) in the model with scores (that is the model $[ART \vee N \bowtie AGR]$), the verb *to_die* does not cluster with any other one, while *to_elapse* heads a cluster with the transitive verb *to_allege*. Thus, there are some clustering errors even after the introduction of this cue, however all of them seem to be related with the incorrect identification of the preposition "a" as an article[4]. So, both *to_allege* (clustered with *to_elapse*) and *to_have* and *to_continue* (heading, respectively, the clusters of *to_be_born* and *to_elapse*), accept as argument a prepositional phrase headed by Portuguese preposition *a*. Probably, if we want to continue to use the LUSA corpus (the only one available and already processed with a dimension sufficient for this kind of analysis), only with the use of more informed and necessary attributes (resulting, for example, from the automatic correction of the occurrences of this type of errors (Rocio Lopes 1999), by using a partial parser (Rocio Lopes 1998) or other opportunistic techniques for partial analysis (Argamon et al. 1998), (Daelemans et al. 1999), will it be possible to decide about this problem.

Finally, both *to_lack* and *to_concur* are correctly clustered together. We still need to refer the fact that the verb *to_depend* has been excluded due to the low number of occurrences of the attribute *n.agr*. As we previously said, in these cases, we should assume the classification given by the model $[ART, AGR]$ (in this in case, the correct subcategorization would have been assigned to the verb).

## 7 Conclusions

In the experiments presented in this paper we have shown how the conjunction of some different attributes in the same loglinear model can improve the acquired results. The use of interactions between several attributes is essential when we intend to make more elaborated studies on subcategorization in general or verbal subcategorization in particular. Many of the traced errors can be easily solved through the consideration of more attributes, as it was the

[4]The used *corpus* was based on the Portuguese News Agency (Lusa) texts. Since the texts are mainly intended for the news professionals and should arrive as earlier as possible, they have lots of errors. These errors are not detected by the part-of-speech tagger. The present error was due to the lack of the diacritic in the contraction of Portuguese feminine singular article *a* with preposition *a (to)* (this gives rise to the Portuguese contraction à).

example of studied categorical variable $AGR$, or many others that the dimension and the focus of the work done until now didn't allow us to study. For example, the simple use of an attribute as *non_ambiguous_article* (that is, all Portuguese articles other than the feminine singular article) can avoid a lot of tagging errors. The number of possible cues for subcategorization learning is probably infinite, and it is more an art than a science to find the right ones (although attribute selection techniques can help, they don't solve the problem). It wasn't our goal to select the right attributes but only to show how to model some of the available attributes. We think, that we have demonstrated that diverse attributes, can be integrated in one same categorical variable, or under diverse categorical variables and then modeled by a loglinear model. Being exponential models, loglinear models are also promising, since it can be demonstrated that, in the case of exponential models, the maximum entropy parameter estimation is equivalent to the maximum likelihood parameter estimation (Ratnaparkhi 1998).

We should also notice that, according to achieved results, it can be observed that the inclusion of more data in the modeling process must be followed by an increase of the descriptive capacity of the loglinear model used. Thus, as in any another modeling task, the maximum of possible information on the behavior of the attributes used must be enclosed in the statistical model. However, only after a thorough analysis of the data and understanding of the meaning of the used parameters, we can add the refinements that make possible to build models that satisfactorily describe the observed data.

As it was experimentally verified in the model of simple independence, the inclusion of too many attributes in the clustering process, demands the corresponding increase in the descriptive capacity of the used model. Thus, it is necessary to maintain a good balance between the modeling capacity of the system and the number of attributes to describe. So, only the attributes that are necessary for the study of the subcategorization class in cause must be used, or else the number of clusters can become extreme, and no advantage for the use of the modeling will be achieved. Equivalently, it is up to the loglinear model to get the best possible description of the observed data. Thus, we must oppose to what (Franz 1996) declares: not always the increase of the number attributes increases the descriptive power of the model (or, in the case of (Franz 1996), making to correspond more interactions to an increase in the number of variables of a logit model ((Franz 1996), (Agresti 1990)), may not correspond to an improvement in the results foreseen for the model: the introduction of more information is not always beneficial. Another of the problems with the fact of considering more attributes is the increasing division and soon reduction of the number of counts for each cell of a given contingency table. Thus for some verbs it is not possible to use more than a complementary pair of attributes. To be able to cope with these cases, the use of more complex models will have to be preceded by the previous use of a simpler model (as it was the case of the model $[ART]$).

The construction of loglinear models on a corpus of texts seems therefore to be a good way to describe the subcategorization classes for a set of verbs. The use of attributes with some relevance to the study being done is needed and it is also needed to use more descriptive models than the simple model of independence. Only then it becomes possible to use the full universe of attributes and models necessary for the characterisation of the different classes of subcategorization.

## 8 Acknowledgements

The work reported in this paper was done in the framework of projects DIXIT[5] and PGR[6]. Nuno Marques was funded by Ph.D. scholarship BD-2909, in the framework of PRAXIS XXI programme.

## References

S. Argamon, I. Dagan, and Y. Krymolovski. A memory-based approach to learning shallow natural language patterns. In *Proceedings of the 36st Annual Meeting of ACL*, pages 67–73, Montreal, 1998.

[5]contract number 2/2.1/TIT/1670/95, funded by programme PRAXIS XXI by the Portuguese Research Funding Agency, JNICT, recently renamed as Fundação para a Ciência e Tecnologia

[6]contract number LO59-P31B-02/97, funded in the framework of programme PRAXIS XXI, measure 3.1-b for consorcia research, by Agência de Inovação

Alan Agresti. *Categorical Data Analysis*. John Wiley and Sons, 1990.

Ted Briscoe and John Carroll. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 356–363, 1997. URL: http://xxx.lanl.gov/ps/cmp-lg/9702002.

Michael R. Brent. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computacional Linguistics*, 19(2):245–262, 1993.

Winfried Busse. *Dicionário Sintáctico de Verbos Portugueses*. Livraria Almedina, 1994.

W. Daelemans, S. Buchholz, and J. Veenstra. Memory-based shallow parsing. In *Proceedings of the EACL Workshop on Computational Natural Language Learning (CoNLL99)*, pages 53–60, Bergen, Norway, 1999.

Alexander Franz. *Automatic Ambiguity Resolution in Natural Language Processing*, volume 1171 of *LNAI Series*. Springer, 1996.

M. J. R. Healy. *GLIM: An Introduction*. Clarendon Press, Oxford, 1988.

Cristopher Manning. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of ACL*, pages 235–242, 1993.

Nuno C. Marques. *Uma Metodologia Para a Modelação Estatística da Subcategorização Verbal*. PhD thesis, Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia, 2000. In Portuguese.

N.M.C. Marques, J.G.P. Lopes, and C. A. Coelho. *Learning Verbal Transitivity Using Loglinear Models*. In Lecture Notes in AI (LNAI): Proceedings of the 10th European Conference on Machine Learning, Claire Nédellec and Céline Rouveirol eds. Springer Verlag, Berlin, April 1998.

N.M.C. Marques, J.G.P. Lopes, and C. A. Coelho. *Using Loglinear Clustering for Subcategorization Identification*, pages 379–387. In Lecture Notes in AI (LNAI): Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery, Jan M. Zytkow and Mohamed Quafafou eds. Springer Verlag, September 1998.

Adwait Ratnaparkhi. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. PhD thesis, University of Pennsylvania, 1998.

Vitor Jorge Rocio and José Gabriel P. Lopes. Partial parsing, deduction and tabling. In Bernard Lang, editor, *Actes des premières Journées sur la Tabulation en Analyse Syntaxique et Déduction (Proceedingsof the Workshop on Tabulation in Parsing and Deduction)*, pages 52–61, Rocquencourt, France, April 2-3 1998. INRIA.

Vitor Jorge Rocio and José Gabriel P. Lopes. An infra-structure for diagnosing causes for partially parsed natural language input. In *ACTAS-I VI Simposio Internacional de Comunicación Social (Proceedings of the 6th International Symposium on Social Communication. Santiago de Cuba, Ja*, pages 550–554, Santiago de Cuba, 1999. Editorial Oriente. ISBN 959-11-0250-X.

Emídio Silva and António Tavares. *Dicionário dos Verbos Portugueses*. Porto Editora, 1989.

Helena Ventura and Maunela Caseiro. *Dicionário Prático de Verbos Seguidos de Preposições*. Fim de Século Edições, LDA., 2 edition, 1992.