Procesamiento del Lenguaje Natural, Revista nº 19, Septiembre de 1996

A Formal Approach to Spanish Morphology: the COES Tools

Santiago Rodríguez, Jesús Carretero
Facultad de Informática
Universidad Politécnica de Madrid (UPM), Spain
Fax: +34-1-3367373
e-mail: srodri@fi.upm.es

Abstract

This paper presents COES, a complete environment which allows the user to deal with Spanish morphological problems. Special emphasis has been made on formal specification of Spanish morphology, word tagging and dictionary generation. Finally, some evaluation results of the spelling services are shown. The main task of this work has been to formalize the set of Spanish rules, which is very complex due to the high number of rules. COES has been integrated with the *ispell* software tool and it is being distributed under the terms of the Free Software Foundation "General Public License" since the end of 1994.

keywords: Computational Linguistics; Spanish; Automatic Spell Checking; Formal specification.

1 Introduction

The increasing use of computers in language processing has shown the lack of some specialized tools (spelling checkers, grammatical checkers, etc.) for the Spanish Language. Spanish is the third most extended language of the world [3]. However, the availability of linguistic tools is far from that of other less extended languages, perhaps due to the smaller technological influence of the Spanish-speaking countries. The potential importance of the Spanish Language in a near future led the authors to develop some lexical tools for computer environments. Moreover, to enhance the utilization of these tools, we thought to distribute them freely with the *ispell* tool. [1] [11].

Four main objectives were established for COES. It should be *exhaustive*, including most of the morphological rules of the Spanish language. It should be *dynamic*, to add new improvements and suggestions. It should be *flexible*, to allow the users to customize the tool. This feature is very important due to the existence of several regional variants of the Spanish language, especially in America. Moreover, to promote its utilization a main goal for COES was to be a freely distributed application.

The main problem found while building this tool was to formalize the Spanish morphological rules, which, opposite to the English ones, are complex and numerous. Thus, to formalize the Spanish morphology *derivation rules* and to obtain a tagged *basic lexicon* were the main tasks in COES tool development. The COES project started at the beginning of 1994. The first prototype of the tool was ready for internal use by mid 1994 and it is being freely distributed since December 1994 [2].

[1] Ispell tool can be obtained by anonymous ftp from `ftp.math.orst.edu` in `/pub/ispell-3.1/ispell-3.1.20.tar.gz`

[2] Anonymous ftp from `ftp.fi.upm.es` in `pub/unix/espa~nol.tar.gz` or by using the http address `http://www.datsi.fi.upm.es/~coes/`

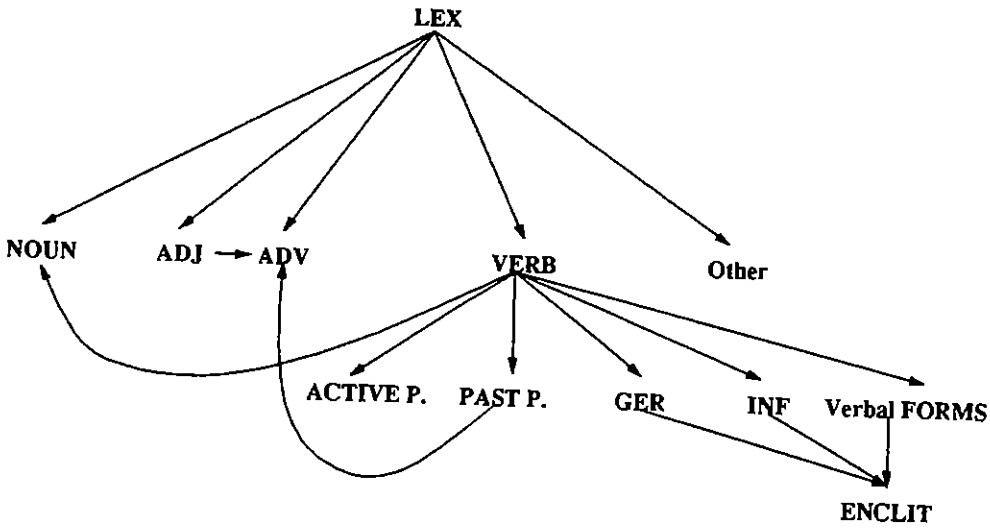

Figure 1: Structure of the Spanish Words

2 Spanish Morphological Features

Spanish is a Latin derived language with a very complex grammar [4, 9]. A simplified version of the tree of morphological features of the Spanish words is show in figure 1. This tree was used to extract the derivation rules of the Spanish Morphology. Due to the intrinsic features of the language, many problems were detected to build the derivation rules, namely:

Gender and number derivations. The *regular* adjectives (ADJ) and substantives (NOUN) have gender (male or female) and number (singular and plural). An example is the noun *perro* (dog), which has three more derivations: *perra, perros, perras.* However, some *irregular* words have only gender (e.j. *casa, casas* (house)), and others have only singular form, but neither gender nor number.

Verb conjugation. There are three different verb conjugations in Spanish depending on the endings (*-ar, -er, -ir*) of its infinitive form. Each conjugation has more than 40 temporal derivations (ACTIVE P., PAST P. GER, INF and Verbal FORMS). In addition, Spanish has two types of verbs: *regular*, which the same derivation rules for all the conjugation, and *irregular*, which have at least one different derivation from the regular verbs. Fortunately, even the irregularities are classified in almost 100 different cases [4].

Enclitic forms. Some verb derivations are generated by adding a pronominal form at the end of a verbal form (ENCLIT). Two different kinds of enclitics are found in written Spanish, depending on the suffixes added to the infinitive and gerund forms: *pronominals*, generated with suffixes *-me, -te, -se, -nos* and *-os*, and *transitives*, generated with *-lo, -la, -los, -las, -le* and *-les*. Some examples are: *amar* → *amarte*, *amar* → *amarse*, and *amar* → *amarse*. The two forms can be combined together (*ajustar* → *ajustármelo*), which generate sets of rules of complexity $O(2)$ and $O(3)$ for the enclitics.

Nouns derived from verbs. Some nouns are derived from a verb such as *imaginar* → *imaginación* or *abatir* → *abatimiento* (VERB → NOUN).

Adverbs derived from adjectives. Many modal adverbs are generated by adding the suffix *-mente* to an adjective (ADJ → ADV): *tranquilo* → *tranquilamente*.

Superlatives and diminutives. Regular superlatives are formed by adding the *-ísimo* to an adjective (*grande* → *grandísimo*). Diminutives are formed by adding the *-ico, -ito* and *-illo* to an adjective or noun.

Acute characters. There are many particularities related with gender and number. words lose its accent mark and change it by the non-acute one: *gañán, gañanes.*

3 Formal Model

A formal model of the Spanish morhology has been defined and implemented relying on sets:
[$\mathcal{X}$.] Lexicon. Unique words that will be used to generate the inflected forms.
[$\mathcal{L}$.] Lexemas.
[$\mathcal{M}$.] Morphemes used, with the $\mathcal{L}$ and $\mathcal{X}$ sets, to generate the inflected forms.
[$\mathcal{W}$.] Set of correct Spanish words, including all the inflected forms.

All the lexicon entries are coded with a predicate that corresponds to its morphological category. The inventory of some categories follows:

- Nouns and adjectives lexemes (`NAL`).
- Regular verb lexemes (`RVL`).
- Irregular verb lexemes (`IVL`).
- Regular (`V`) and irregular (`W`) verb morphemes.
- Nominal gender and number morphemes (`P` and `S`). `P` is the set of morphemes belonging to the gender and number category and `S` is only for number.
- Gerund and past participle morphemes (`X` for regular verbs and `Y` for irregular ones).
- Enclitic pronominal morphemes (`R` for regular verbs and `O` for irregular ones).
- Enclitic transitive morphemes (`T` for regular verbs and `Q` for irregular ones).
- Morphemes to generate adverbs derived from adjectives (`M`).

Such categories are modeled with macrorules, which reflect one particular aspect of the Spanish morphology. As Spanish strongly relies on inflected forms (a Spanish verb has more than 55 inflected forms), our derivation rules include all the *regular* inflectional behavior of the Spanish words. Moreover, we have implemented extra rules to capture the regular patterns of the irregular verbs and to include singular words not belonging to the previous models (verbs *ser, ir, haber*, and *estar*, some number and gender irregularities as *caballo* and *yegua*, etc.) [15]. Almost 3,500 rules have been used in COES to specify the Spanish morphology.

The lexemes are built using the generic method *lex* which extracts the lexeme of a word by applying every morpheme in the second argument. If a valid morpheme is found, the lexeme is obtained by taking the morpheme out of the word.

```
lex(lemma, <morpheme list>)
```

Some examples of categories usage, word tagging and inflected form generation are shown in the following paragraphs: **Gender and number derivations.** Some morphemes for gender and number for the P macrorule are:

```
P(pl1,  [masc, fem], plural) -> S      P(pl2,  masc,       plural) -> ES
P(pl4,  fem,         plural) -> AS     P(pl3,  masc,       plural) -> CES
P(gen1, masc,          sing) -> 0      P(gen2, fem,          sing) -> A
P(gen3, masc,          sing) -> E
```

Some examples of P macrorule application to the Spanish words *pastor* (shepherd), and *presidente* (president) are shown below:

```
lex(pastora, [gen2,pl2,pl4])     -> pastor
gen2(pastor)->pastora
pl4(pastor)->pastoras
lex(presidente, [gen2,pl1,pl4]) -> president
gen2(presidente)->presidenta
pl4 (presidente)->presidentas
```

Verb conjugations have been defined using 4 macrorules which describes regular and irregular verbs. Around 200 rules compose the regular verbs derivations and 2,700 the irregular ones. Irregular verb formalization was considered very important because Spanish has a lot of irregular forms in its morphology that follow a well defined derivation pattern (*-ontar* → *-uento*, *-oder* → *-uedo*, *-ervir* → *-irvo*, etc). Excluded from these rules are the verbs *ser*, *estar*, *ir* and *haber* because no way to derive the different forms of these verbs from the infinitive has been found. Instead, all their forms have been explicitly included in the lexicon. The specification of some verb morphemes is shown below. The argument *ZC* used for the *W* category specifies the type of irregularity to be applied.

```
V(inpr1s,conjugation,indicative,present,1st.person,sing)       -> O
V(rinpr2s,conjugation,indicative,present,2nd.person,sing)      -> AS
V(rinpa2s,conjugation,indicative,past,2nd.person,sing)         -> ASTE
V(rsufu2p,conjugation,subjunctive,present,1st.person,plur)     -> EMOS
V(rsufu2s,conjugation,subjunctive,future,2nd.person,sing)      -> ARES
W(iinpr1s,conjugation,ZC,indicative,present,1st.person,sing)  -> OZCO
W(isupr3s,conjugation,ZC,subjunctive,present,3nd.person,sing) -> OZCA
```

Some derivation rules for the regular verb *amar* (to love) and for the irregular one *conocer* (to know) are:

```
rinpr1s(amar,[ar]) -> amo
rinpr2s(amar,[ar]) -> amas
rsufu2s(amar,[ar]) -> amares
iinpr1s(conocer,[er]) -> conozco
isupr3s(conocer,[er]) -> conozca
```

Some new classes have been defined to generate **enclitic forms** and **nouns derived from verbs**. The enclitics of regular verbs require around 200 derivation rules and the irregular ones around 450. These rules specify the behavior of pronominal, transitive, and combined derivations. All the rules are applied only to infinitive and gerund forms. Nouns derived from verbs are generated using two classes: nouns ending in *-miento* and *-ción*. Derivations from adjectives include **adverbs, superlatives, and diminutives**. Derivations generated by adding prefixes can not be grouped. Each one has been defined by a macrorule, resulting in a set of 20 macrorules for the most commonly used.

The model described has been adapted to the *ispell* formalization requirements, which follows a finite-state processors model [13, 18]. Some constraints on the syntax of the rules and classes led to the replication of some rules in several classes. To show how the former rules are applied, a simplified schema of the inflected forms of the word *deber* is shown in figure 2.

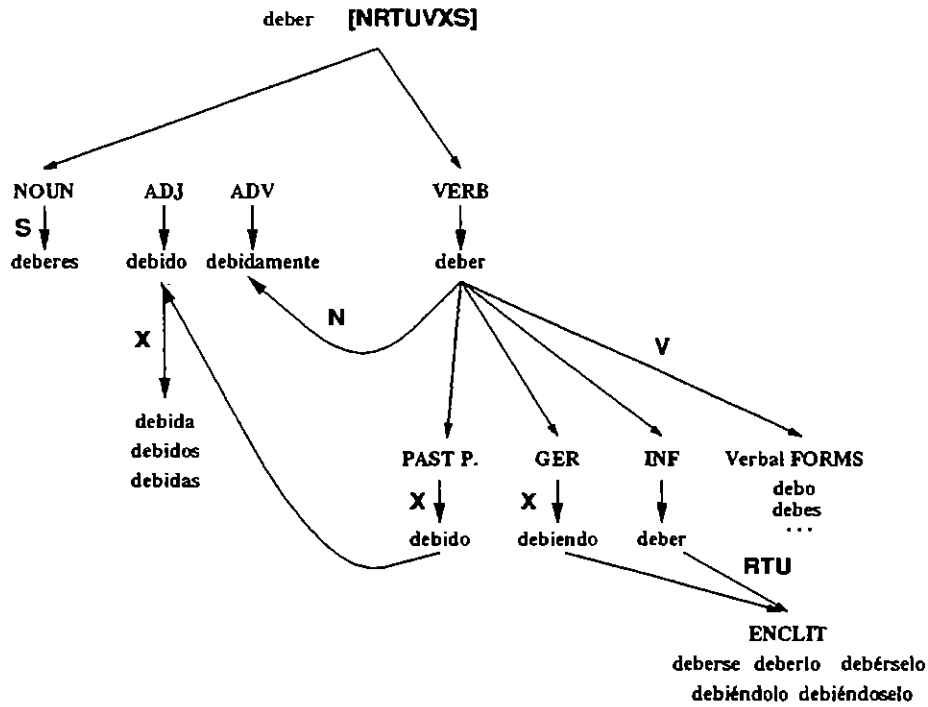

Figure 2: Inflected Forms of *deber*

4 The Processor

In order to define the morphological relationship between the former sets, each entry of the lexicons is tagged with rules. These rules allow to generate the $\mathcal{W}$ set from the $\mathcal{X}$ and $\mathcal{M}$ sets. A finite-state processor approach [14, 18], implemented using a set of functions defined below, has been used to generate and recognize Spanish words.

The function *islex* performs the pattern matching between the morphemes and the words of the lexicon to extract valid lexemas. It is applied to the list of valid morphemes of each lexeme, which should be described with a `lex` rule.

$$\forall x \in \mathcal{M},\ \forall y \in \mathcal{X}\ ,\ \exists islex(x,y) : \mathcal{M}\ \mathrm{x}\ \mathcal{X} \rightarrow \{0,1\}$$

$$islex(x,y) = \begin{cases} 1 & \text{if } y - x \in \mathcal{L} \\ 0 & \text{otherwise} \end{cases}$$

The function *ismor* performs the pattern matching between the morphemes and the Spanish words. If a Spanish word matches a morpheme and the resulting string is recognized as a lexeme, this function is used to trigger the reduction rules (R).

$$\forall x \in \mathcal{M},\ \forall y \in \mathcal{W}\ ,\ \exists ismor(x,y) : \mathcal{M}\ \mathrm{x}\ \mathcal{W} \rightarrow \{0,1\}$$

$$ismor(x,y) = \begin{cases} 1 & \text{if } y - x \in \mathcal{L},\mathcal{W} \\ 0 & \text{otherwise} \end{cases}$$

An *expansion* rule, $E(x, y, z)$, is an application to obtain an inflected form from a lexicon entry. Each rule is triggered when the *islex* function finds a valid lexeme by using the first parameter of the function (x) and the lexicon entry (y). The result of the derivation rule is a Spanish word built by adding z to $y - x$ (lexeme).

$\forall x, z \in \mathcal{M},\ \forall y \in \mathcal{X}\ , \exists E(x, y, z) : \mathcal{M}\ \mathrm{x}\ \mathcal{X}\ \mathrm{x}\ \mathcal{M} \rightarrow \mathcal{W}$

$$E(x, y, z) = \begin{cases} y - x + z & \text{if } islex(x, y) = 1 \\ \emptyset & \text{otherwise} \end{cases}$$

A morphological class is defined as a set of rules inherited by all the members of the class. Formally a class is defined as a set of rules (E_i) to be applied to a word belonging to the lexicon. This set of rules always include the $E(\emptyset, x, \emptyset) = x$ rule.

$\forall y \in \mathcal{X}\ , \forall i = 1, \ldots n, \quad x_i, z_i \in \mathcal{M}, \exists C(y) : \mathcal{X} \rightarrow \mathcal{W}$

$$C(y) = \bigcup_{i=1}^{n} E_i(x_i, y, z_i) \cup E(\emptyset, y, \emptyset)$$

As shown in the previous paragraphs, this model works by doing string pattern matching and concatenation on the components of the lexicon and morpheme sets. In order to use this model, the existence of a *tagged* lexicon is needed. A tag is a list of classes whose rules may be applied to the lexicon word. Building a complete dictionary can be done by applying every rule of every class to the tagged lexicon.

A *reduction* rule, $R(x, y, z)$, is an application to obtain a root word from an inflected form. Each rule is triggered when the *ismor* function finds a valid morpheme by using the last parameter of the function (z) and the Spanish word (y).

$\forall x, z \in \mathcal{M},\ \forall y \in \mathcal{W}\ , \exists R(x, y, z) : \mathcal{M}\ \mathrm{x}\ \mathcal{W}\ \mathrm{x}\ \mathcal{M} \rightarrow \mathcal{X}$

$$R(x, y, z) = \begin{cases} y - z + x & \text{if } ismor(z, y) = 1 \\ \emptyset & \text{otherwise} \end{cases}$$

Word reduction is accomplished by using a tree adjoining methodology [10, 1] with a finite set of elementary trees, each of which is a domain of locality, and can be viewed as a minimal morphological structure. The trees used by COES are derived from the general one shown in figure 1. The initial trees show the main morphological classes (noun, verb, adjective, etc.) and their derivations. The auxiliary trees show connections among classes. Our reduction operation is similar to the adjunct, but applied reversely to go from the leaves to the root of the trees. Applying reduction successively, valid morphemes at an interior node can be extracted from a complete structure to get a reduced structure which can be checked with the ones found in the lexicon ($\mathcal{L}$).

A Spanish word is recognized in COES by successively reducing the valid morphemes found in the word. If the result is found in the lexicon, the word is recognized as correct. Otherwise, it is not recognized and it is pointed out as erroneous. To optimize the word recognition process, the reduction rules are only applied when the word is not found in the dictionary generated by expanding the lexicon entries.

$\forall y \in \mathcal{W}\ ,\ y\ recognized \rightarrow \exists x_0 \ldots x_n, z_0 \ldots z_n \in \mathcal{M},\ R(x_0, \ldots R(x_n, y, z_n) \ldots, z_0) \in \mathcal{L}$

5 The Tagset and the Lexicon

The lexicon for this platform has been extracted from a *Spanish corpus* compiled by the aut
This corpus, including more than 20 million words, comprises texts extracted from Spa
newspapers (ABC Cultural, El Mundo, El Periódico, etc.), selected books, technical bo
from the UPM library, oral corpus [12], and the concise version of the Collins English-Spa
Dictionary. The main criteria used to chose the lexicon was to represent the main class
Spanish words, representing as many set of ambiguities as possible [17]. Of course, categ
that are too small were avoided. The *basic lexicon* used as starting point in COES inclu
currently more than 80,000 different words.

To generate a dictionary, the lexicon must be tagged according to the derivation rules
the COES formal model described above. As manual tagging is a very hard and error pr
work, some tools have been developed in COES to tag the lexicon entries semiautomatical
extracting the morphemes of each word and proposing a *tentative* tag for them. This tag
has been trained using the *bootstrapping* method [6]. A small manually tagged lexicon, wh
correctness was guaranteed, was used to train the tagger. Then, the tagger was used to t
more lexicon words, which were partially corrected and fed to the tagger trainer. This meth
is highly adequate for inflectional languages, like Spanish, where there is a clear correspond
between suffixes and morphosyntactic properties of the words. Some other tools [2, 16] use t
former model combined with statistical ones.

As COES is designed to analyze large archives, the words not found in the lexicon a
analyzed by a separate finite-state machine which is very efficient and compact. It is highl
improved to detect the main failure causes in COES (pronominal verbs, enclitics and neol
gisms), because, fortunately, most of them follow a regular inflectional pattern. Tentative ta
are made based on *productive endings* [2] very usual in Spanish: *-mente* for adverbs, *-ura* fo
adjectives, *-ante* for nouns, etc. These endings were extracted by computing the most frequen
ending patterns in the corpus, but discarding highly frequent words with irregular endings. T
be accepted, the new words and their tags must be manually checked by looking in the *referen
lexicon* [5].

6 Conclusions and Future Work

A Spanish dictionary for *ispell* has been developed and it is being used by a large community
Our feeling is that the dictionary works properly and it is very exhaustive. The error rat
related to the corpus size is approximately 0.4 % of the words present in the corpus. However,
the number of unrecognized words over the list of unique words present in the corpus is around
2.5 %. This error rate is mainly due to prefixes, enclitics, comparatives, and local expressions.
COES average performance is around 1,600 $\frac{words}{second}$, a very good result compared to other spell
checkers (e.j.: 5 $\frac{words}{second}$ in Aries [7]).

COES tool is being improved in the following aspects: elaboration of geographical and
thematic dictionaries, optimization of rules, and lexicon improvement. The proposal is to
create lexicons and dictionaries per geographical areas (Chile, Perú, Colombia, etc.) Moreover,
some thematic lexicons for computer science, legal or medical terminology are now in progress.

Some new COES utilities are being developed. A *thesaurus*, which extensively uses the
derivation rules, will be available soon. A preliminary study of Spanish syntax rules and
Spanish morphological models [8] is being developed. The purpose is to build an efficient
Spanish syntax analyzer using the TAG model to represent the sentences.

References

[1] A. Abeillé. A French Tree Adjoining Grammar. Technical report, Univ. of Pennsylvania, Philadelphia, USA, 1988.

[2] J. Chanod and P. Tapanainen. Tagging French: comparing a statistical and a constraint-based method. In *Proceedings of the EACL-95*, 1995.

[3] B. Comrie. *The World's Major Languages.* Croom Helm, London, 1987.

[4] Real Academia Española de la Lengua. *Esbozo de una Nueva Gramática de la Lengua Española.* Espasa Calpe, 1991.

[5] Real Academia Española de la Lengua. *Diccionario de la Lengua Española.* Espasa Calpe, 21 edition, 1992.

[6] A. Derouault and B. Merialdo. Natural Language Modelling for Phoneme-to-Text Transcription. In *IEEE Transactions on Pattern Analysis and Machine Intelligence , PAMI-8*, pages 742–749, 1986.

[7] J. González, J.M. Goñi, and A. Nieto. Aries: a ready for use platform for engineering spanish-processing tools. In *Digest of the Second Language Engineering Convention*, pages 219–226, October 1995.

[8] J. Hallebeek. A Formal Approach to Spanish Grammar. *Language and Computers: Studies in Practical Linguistics*, 1992.

[9] J.Hallebeek. *Morfología y Sintaxis del Español: Introducción al Análisis Oracional.* Playor, Madrid, Spain, 1994.

[10] A. Joshi. Tree Adjunct Grammars. *Journal of the Computer and System Sciences*, 10(1):136–163, 1975.

[11] G. Kuenning. Ispell tool. Enclosed with the ispell distribution, 1995.

[12] F. Marcos, A. Ballester, C. Santamaría, E. Pertierra, O. Brandeo, and P. Díez. Corpus oral de referencia de la lengua española comtemporánea. Technical report, Universidad Autónoma de Madrid, 1992.

[13] M. Meya. Morphological analysis of Spanish for retrieval. *Literary & Linguistic Computing*, 2(3):166–170, 1987.

[14] A. Moreno and J. Goñi. GRAMPAL: A Morphological Processor for Spanish implemented in Prolog. Technical Report TIC91-0217C02-01, CYCIT, Madrid, Spain, 1995.

[15] S. Rodríguez and J. Carretero. Building a spanish speller. In *Taller sobre Software de Libre Distribución.* Universidad Carlos III de Madrid, Spain, 1995.

[16] F. Sánchez and A. Nieto. Development of a Spanish Version of the Xerox Tagger. Technical Report CRATER/WP6/FR1, CRATER Project, CEC, 1995.

[17] P. Tapanainen and A. Voutilainen. Tagging accurately – Don't guess if you know. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, August 1994.

[18] E. Tzoukermann and M. Liberman. A Finite-State Morphological Processor for Spa
*Proceedings of the 13th International Conference on Computational Linguistics (CO
90)*, pages 277–281, 1990.