TOWARDS AN INTEGRATED ENVIRONMENT FOR SPANISH DOCUMENT VERIFICATION AND COMPOSITION

R. Casajuana, C. Rodríguez, L. Sopeña, C. Villar
IBM Madrid Scientific Center
Paseo de la Castellana, 4
28046 Madrid

ABSTRACT

Languages other than English have received little attention as far as the application of natural language processing techniques to text composition is concerned. The present paper describes briefly work under development aiming at the design of an integrated environment for the construction and verification of documents written in Spanish. In a first phase, a dictionary of Spanish has been implemented, together with a synonym dictionary. The main features of both dictionaries will be summarised, and how they are applied in an environment for document verification and composition.

INTRODUCTION

In the field of document processing many tools exist today which allow the user to introduce a text in storage, format it, and even, for a few languages, verify the spelling, punctuation and style [1, 2, 3, 4]. English has been for a long time THE Natural Language, object of a large number of research and development work in Computational Linguistics. Other languages, however (Spanish among them), have received little attention as far as the application of natural language processing techniques to text composition is concerned.

The present paper describes briefly work under development aiming at the design of an integrated environment for the construction and verification of documents written in Spanish, for which no similar tools exist at the moment.

In a first phase, a dictionary of Spanish was implemented. This is a task of multiple interest, a dictionary being the one of the basic tools for any application to systems where Natural Language is involved. Thus its development was undertaken with two guidelines, completeness and generality. At present, the dictionary is finished in a version including about 35,000 stems, which, inflected, give rise to more than 400,000 different words.

Together with this inflected forms lexicon, a synonym dictionary was also built as a second step in the text processing system; this dictionary has about 15,000 entries.

In this paper we summarise the main features of both dictionaries and how they are applied in an environment for document verification and composition. Present and planned enhancements will be also described, including the use of a parser of Spanish and the addition of other features.

THE INFLECTED FORMS DICTIONARY

The starting point was an analysis of word frequency performed on different texts previously selected: press articles, novels, essays, etc. totalling approximately one million words. A listing of the whole set of the entries of the **Diccionario de la Real Academia Española** [5] (DRAE, Dictionary of the Spanish Royal Academy, containing the "official" Spanish language) was studied, and several other published dictionaries were as well collated [6, 7, 8, 9]. The information so obtained was classified and filtered, taking into account the objective and first set up application: the corpus had to cover *usual written language*, and in this field should account for as much of the vocabulary as possible.

The dictionary consists of a list of inflected words, without associated definitions. Every word has additionally a number of other information: gender, number, time, person, mode, etc.

In general, words belonging to restricted or specialised domains (medicine, law, poetry, linguistics, etc.) are not listed. Neither are colloquial terms, including rude or slang words. Very specific regional uses of Spanish have also not been considered (like Argentina's "voseo": *tenés, querés*), nor the form of subjunctive future (*tuviere, quisiere*), restricted today to legal writings. Many derived forms have also been excluded, like diminutives, pejoratives, superlatives (but not the irregulars); as for adverbs finishing in **-mente**, only the most usual ones have been listed.

Information on the lexicon is contained in two main files: the base forms file, and the inflective morphemes file, which are described in the following sections.

Base forms file

It includes the complete list of terms just described, specifying the base form on which they inflect. They have pointers referring to the derivative morphemes file.

Each entry has the following specifications:

1. Functional category, i.e., verb, noun, adjective, adverb, preposition, conjunction, article, pronoun, interjection; words with more than one

associated part of speech will have as many marks as categories.

2. Verbs, very complex because of the large number of irregularities and difficult classification, are qualified as transitive, intransitive or auxiliary. Further slots are foreseen to code their behaviour in the language and their usage at the surface level: complements, adverbials, etc. Possible combinations of verbs and clitic pronouns are also marked.

3. There are additional marks for hyphenation points (for later use by a formatter performing automatic syllable partition), and several other for foreign and Latin words, geographical terms, etc.

Inflective morphemes file

It specifies the derivative morphemes used in the generation of inflected forms starting from the previous base forms. A list of paradigms has been built for each category of nouns, adjectives and verbs, to account for the different models of inflection.

The classification takes into account the problems arising from the automatic processing of inflections, i.e., it considers as irregularities some behaviours not considered as such in the literature, for example, some purely phonetic cases, like z → c before e, i (e.g. *cazar → cace*), and cases related with diacritic signs, both dieresis (e.g. *avergonzar → avergüenzo*), and accents (e.g. *joven → jóvenes, carácter → caracteres*).

Additionally, it is necessary to consider cases of incomplete inflections (e.g. in adjectives, *avizor* only exists in masculine singular, and *alisios* only in masculine plural; in names, *alicates* exists only in masculine plural, *afueras* only in feminine plural). As for verbs, this kind of irregularity is present in the so-called defectives (e.g. *llover, abolir, pudrir, etc.*). Finally, there are words with more than one realisation in one of their forms (e.g. *variz/varice*, both correct in feminine singular). In some adjectives, a similar problem arises depending on their position: if they come in front of the noun their apocopated form appears, but not if they come after (e.g. *buen/bueno, mal/malo*), and in verbs, in all subjunctive imperfect forms (e.g. *saliera/saliese*), and in a few other isolated cases (e.g. the imperative *satisfaz/satisface*).

Together with adjectives marked for gender (e.g. *rojo, roja*), there are others unmarked (e.g. *amable*), and their gender is defined according to the noun they modify. Among them, some work in fixed and restricted contexts, and are defined because they only modify masculine or feminine nouns (e.g. *torcaz, avizor*).

It must be noted that the large number of irregularities in the inflection mechanism has obliged to detail each one of them, as they could not be included in any of the general models. This means that many paradigms have been defined which just comprise a little number of cases. The complete description of the classification performed has been the object of previous papers [10, 11].

THE SYNONYM DICTIONARY

To build the synonym lexicon, a published dictionary was used [12], which had to be modified due both to the specific needs of computer processing and to the many typographical errors and inconsistencies found in its contents. This has allowed to develop a thorough study on synonymy together with a complete critique of one of the best-known synonym dictionaries of the Spanish language.

First of all, the coherence of both dictionaries has been kept, so that words included in the synonym base are also present in the main lexicon.

The need to keep the semantic consistency in the dictionary contents was a first objective. It showed the little rigor with which printed dictionaries are constructed and allowed for the application of systematic tests and modifications to our version in order to keep symmetry, to cater for hyperonymy, to bind cross-referencing into semantically reasonable limits, etc. A forthcoming paper will describe the problems met and the main tasks performed.

Starting from syntactic marks in the inflected forms dictionary, an entry in the synonym dictionary will appear as many times as parts of speech it is assigned. For example, the word *circular* can be an adjective (marked as j, meaning 'circular'), a feminine noun (marked as nf, meaning 'note'), and a verb (marked as v, meaning 'move', 'circulate'). The corresponding entries would be:

circular: j
redondo, curvo, curvado.
circular: nf
orden, aviso*, notificación, carta, nota.
circular: v
andar, moverse, transitar*, pasear, deambular;
divulgarse, propagarse, expandirse, difundirse.

Additionally, inside a part of speech, synonyms are grouped according to the different semantic sense or nuance. Also allowed are cross references (marked with asterisks * in the file), which link one synonym to another dictionary entry, thus extending the information power of the lexicon.

More specific information about the entries can also be defined by means of the so-called "qualifiers", which introduce further restrictions on the entry word for that meaning to apply. For example, the noun *costa* means 'coast', but in plural it is also used to mean specifically 'costs'. The verb *echar* has several different senses ('throw', 'dismiss', 'emit', etc.), but its reflexive form *echarse* means 'lie down'.

costa: n
playa, litoral, margen, orilla, borde;
<plural>
cargas, desembolso, importe.
echar: v
expulsar, repeler, rechazar, despachar, excluir;
deponer, destituir;
dar, entregar, repartir;
.....
<se>
tenderse, acostarse, tumbarse, arrellanarse.

DICTIONARY-BASED TEXT COMPOSITION

Spelling verification

The approach is based on the identification of all strings in the text which are not present in the dictionary. Verification algorithms isolate each word (token), look for them in the lexicon and point out to the user which ones have not been found (by highlighting them in the screen or using a different colour). A token is thus every sequence of letters separated by delimiters (in Spanish: blank, comma, period, colon, semicolon, hyphen, open and close question and exclamation marks). The size of the dictionary will have several obvious implications: the frequency of correct words that will be rejected, the search time, the amount of storage allocated. A compromise among all these factors and the use of several compaction mechanisms have allowed its size to remain between reasonable limits.

The spelling verification performed at this moment considers each word in the text independently of the rest.

An additional and interesting possibility of the program is that it allows the user to define his/her own dictionary of addenda, where terms not known by the system (proper names, technical or specific words) can be stored.

Spelling correction

Apart from detecting incorrect terms in the text, the program can also propose for each wrong token a list of candidates, words very similar to the token but which are included in the dictionary. This list is presented with the alternative terms sorted in decreasing priority order, depending on the value of a similarity index computed for each word. This "similarity" is determined by an algorithm, and essentially depends on the number of alterations that must be performed on the token to obtain the correct word. Thus it is a function of the relative difference in length between the token and the word, the difference in the character sequence due to any of the most typical error sources (transcription, omission, insertion, substitution), the matching of the last letter, etc.

The user can choose a word in the proposed list, and the system will automatically replace the wrong term with the selected one.

Morphology function

For each word in the text the program is able to produce all its possible base forms and parts of speech (out of context at this first stage). It can also generate the complete set of derived forms for each of those possibilities. This is most interesting in Spanish in the case of unusual inflections, like many irregular and defective verbs, when in doubt about the use of accents, with some special nouns and adjectives, with seldom used terms, etc.

Synonym function

The mechanism is very similar to the one described for alternative terms: when the user asks for synonyms of a given word in the text these are displayed in a window. At present, words with several parts of speech having specific synonyms for each of them get a multiple display of synonyms for all those parts. For example, synonyms to the word *bajo* will be presented in several lists: as a verb (present tense of *bajar*: 'get down'), as a noun ('ground floor'), as an adjective ('low'), as an adverb ('down'), and as a preposition ('under'). This is, of course, an extreme case, but there are many similar examples.

The user may choose one of the synonyms and automatically replace for it the word in the text. In this first phase, the synonym function does not inflect the candidates in the form of the original token. Starting from it, it performs a morphological analysis, finds its stem and looks for the synonyms in the corresponding dictionary. Thus, if the user writes *Juan quiere a María* ('John loves Mary') and requests synonyms for *quiere*, the system will find the base form *querer* ('to love'), and will display, for example, the infinitive *amar*, but not *ama*, which is the corresponding inflected form (third person singular indicative present) of the original verb. Similarly, when asking for synonyms of *niñas* ('girls'), it will give the list of synonyms for *niño* ('boy'), which is its base form according to the defined paradigms.

PARSING AND OTHER ENHANCEMENTS

A dictionary-based text composition facility is of a great help when writing documents, but it is clearly not enough. Our next objective is to implement a parser of Spanish and to integrate it, as a first application, into the existing system. This will have several consequences in the enhancement of its present capabilities and will add new possibilities of verification.

For example, it will allow the processing of multiple-word phrases, compounds and adverbials. It will make possible for the synonym feature to only propose alternatives for a word in the suitable part of speech and exclude all other possibilities according to the context.

It will also allow to overcome some of the limitations of spelling verification as performed now, by taking into account the context; thus, errors due to the use of correct words (i.e., included in the dictionary) in a wrong syntactic environment, will be detected in most cases. The main causes of confusability now unnoticed that will be highlighted are due to three different types of ambiguity:

- Graphical ambiguity: homophone words with a graphic difference in the accent and with different parts of speech (E.g. relative vs. interrogative pronoun: *cuanto/cuánto*, preposition vs. verb: *de/dé*, conditional vs. affirmative conjunction: *si/sí*, etc.).
- Accentuation ambiguities: based upon the accent change inside a group of words, sometimes with a different part of speech associated (E.g. verb vs. noun: *baile/bailé*, verb-noun-adjective vs. verb: *frío/frió*, noun vs. verb vs. verb: *cántara/cantara/cantará*, verb vs. verb: *ame/amé*, etc.).
- Phonetic ambiguities: implied by orthographic problems based on Spanish phonetics (E.g.*asta/hasta, tubo/tuvo*, are phonetically ambiguous; *callado/cayado, contexto/contesto* also in some regions).

Naturally this would only be the most immediate application of the parser, and it must be noted that some of the described ambiguities will need a great deal of semantic knowledge to be resolved; this we are not considering for the moment. Other obvious uses include the detection of agreement errors: inside Noun Phrases (in Spanish its elements must agree in gender and number), between the subject and the verb of a sentence, errors in the use of pronouns (typical misuses are the so-called "leísmo" and "laísmo"), errors in the order of clitic pronouns, etc.

The different elements integrating the system constitute a set of different pieces whose application is of course not bound to document composition: several other objectives are also foreseen for the dictionaries and the parser, a computer-assisted verb conjugation system has already been built for Spanish grammar students, and other ideas include automatic document abstracting, storage and retrieval, inclusion of dictionary definitions and translation into other languages, and document style critiquing.

REFERENCES

[1] André, J.: **Bibliographie analytique sur les "manipulations de textes"**, Technique et Sciences Informatiques, vol. 1, no. 5, 1982.

[2] Larson, J. A., ed.: "Creating, Revising, and Publishing Office Documents" (Chapter 6), in **End User Facilities in the 1980's**, IEEE, New York 1982.

[3] Cherry, L.: **Writing Tools**, IEEE Trans. on Communications, vol. 30, no. 1, January 1982.

[4] Peterson, J.L.: **Computer Programs for Detecting and Correcting Spelling Errors**, Comm. of the ACM, Dec. 1980, vol. 23, no. 12.

[5] Real Academia Española: **Diccionario de la Lengua Española**, vigésima edición, Ed. Espasa-Calpe, Madrid, 1984, 2 vols.

[6] Moliner, M.: **Diccionario de uso del español**, Ed. Gredos, Madrid, 1982.

[7] Casares, J.: **Diccionario ideológico de la Lengua Española**, Ed. Gustavo Gili, Barcelona, 1982.

[8] **Diccionario Anaya de la Lengua**, Ed. Anaya, Madrid 1980.

[9] Seco, M.: **Diccionario de dudas y dificultades de la lengua española**, 9a. ed., Ed. Espasa-Calpe, Madrid 1986.

[10] Casajuana, R., Rodríguez, C.: **Clasificación de los verbos castellanos para un diccionario en ordenador**, Actas 1er. Congreso de Lenguajes Naturales y Lenguajes Formales, Barcelona, octubre 1985.

[11] Casajuana, R., Rodríguez, C.: **Verificación ortográfica en castellano; la realización de un diccionario en ordenador**, Español Actual, no. 44, 1985.

[12] Sáinz de Robles, F.C.: **Diccionario español de sinónimos y antónimos**, Ed. Aguilar, 1984.