Linguist's Intelligent Workplace for Spanish Language*

Grigori Sidorov

Center for Research on Computer Science (CIC),
National Polytechnic Institute (IPN),
Av. Juan de Dios Batiz, s/n, Zacatenco, 07738,
Mexico City, Mexico
sidorov@cic.ipn.mx

Abstract. The purposes of linguist's intelligent workplace is, first, to perform automatic marking of various linguistic phenomena in texts when there is no ambiguity, and, second, in case of ambiguity, to permit a user (linguist) resolving it manually with the least effort. The paper presents analysis of requirements for the linguist's intelligent workplace, i.e., several marking schemes depending of the language level, interface clarity, natural representation of data (texts), multimodality (usage of colors and fonts, audio), integration of additional NLP techniques, and interaction with user on the basis of menus composed of possible linguistic values (marks). Also the workplace for the morphological level of Spanish language is described.

1 Introduction

Computers are excellent assistants of humans in many tasks. During their first appearance, they served basically for calculations and mathematical modeling, while nowadays they are used in many other tasks related to everyday activities of humans like text editing, communications, presentations, etc. Modern computers possess certain intelligence that allows them imitating of human behavior in some specialized problem solving tasks. This ability is known as artificial intelligence. Computers even present certain linguistic skills; for example, they allow checking of orthographical errors and style while texts preparation.

But how do computers affect a work of traditional linguists? Let us first see what the activities of a linguist are in the most general sense of a word. A linguist dedicates himself to analysis of language phenomena, that is, that he should, first, find sufficient material (empirical examples), then apply certain reasoning based on his experience, analogies from the same language or other languages, and, finally, construct a model that describes the phenomenon. Steps two and three (reasoning and modeling) are based on human knowledge and intuition and, thus, cannot be formalized at the

* Work done under partial support of Mexican government (CONACyT and SNI) and National Polytechnic Institute, Mexico. I would like to thank Dr. Raúl Ávila and Dr. Alexander Gelbukh for useful discussions.

Advances in Artificial Intelligence and Computer Science
Research on Computing Science 14, 2005, pp. 173-178

modern stage of computer science, still, step one (collecting examples) has excellent perspectives in applying computers. One of the branches of computational linguistics – corpus linguistics – is devoted to a great extent to this problem [1]. Roughly speaking, traditional corpus linguistics is a science of how to gather sufficient amount of texts and how to mark them for further extraction of examples for various language phenomena in an easy way.

Still, many traditional linguists search examples as they have been doing in a pre-computer era – by reading texts (even if texts are in a computer form) and by marking relevant examples manually. Other possible approach is just relying on intuition and inventing examples, which is very rapid procedure but it does not guarantee that the examples are complete reflection of language usage.

The purpose of this paper is, firstly, to define the requirements for an environment that would help linguists in marking linguistic phenomena in texts, and, secondly, to describe the environment for such marking developed for morphological level of Spanish language. Though the described environment is oriented to Spanish, general requirements are language-independent.

The paper has the following structure. At the beginning, we discuss general requirements to linguist's intelligent workplace; then, we describe the environment for Spanish language; finally, some conclusions are drawn.

2 General Requirements to Linguist's Intelligent Workplace

There are six traditional basic levels in language description: phonetics/phonology, morphology, syntax, semantics, pragmatics, and discourse. Let us remind that the differences between these levels depend basically on the focus of investigation, for example, there are semantic feature at the morphological level, but we focus on the relations between morphemes, etc.

Each level needs its own encoding scheme, for example, at the phonetic level we mark the pronunciation or intonation; at the morphological level grammar categories of words are important; at the syntactic level we detect relations with other words, presence of certain syntactic constructions or function of words in sentences; at the semantic level we choose word senses used in the given context or mark semantic features; at the pragmatic level we mark intentions or presuppositions; at the discourse level, the attention is paid to anaphoric relations or theme-rheme dynamics, etc.

Note that at each level we should treat the most difficult and wide-spread problem in investigation of natural language – the ambiguity. So, linguists need a corresponding tool. Thus, the general purpose of the linguist's intelligent workplace can be formulated as follows: it should allow for processing of all cases without ambiguity automatically and in case of the presence of ambiguity, it should permit manual resolution with the least effort. In this way, texts are marked semiautomatically. When they are marked, then it is possible to make queries for easily collection of the desired language phenomena for further analysis.

Other approach to the linguist's workplace is development of sets of tools [2]. Disadvantage of this approach is lack of interface possibilities. So, only experimented

users can use them. Other consideration is that sets of tools are not oriented to manual processing from the very beginning.

It is obvious that interaction with user in natural language, even restricted one, (see, for example, [8]) in the case of workplace is unnecessary.

Due to the difference between language levels, it is recommended to provide the possibility of switching between different marking schemes depending of the level that is being analyzed. Still, the same text can be marked at several language levels.

It is also desirable to detect and to mark non-trivial cases (contexts), for example, some rare senses of a word can occur very few times, thus, it is important to have the opportunity not to loose them during analysis among the much more frequent senses. It can be done automatically by calculation of similarity of contexts. There are many methods for evaluation of similarity based on machine learning techniques or lexical resources, see, for example, [4].

Another interesting feature is invoking machine learning techniques, for example, for detecting the most probable item in the suggested menu and propagating it to the top of the menu list. Say, during morphological marking it is possible to use trained POS-tagger for detecting the most probable grammar category. Also, it is interesting to train machine learning methods during mark up, because they obtain more and more manually resolved data.

It is preferable to use menu systems for interacting with user because it reduces the risk of errors. The items in the menu shown to the user are linguistic values and his task is to choose the right value.

A user also should always have the possibility to return to his previous decision (at least, the last one) and change it if necessary. It is important because the user can suddenly realize that he has just made a mistake.

During the interaction with a user, it is also desirable that system's next step would be predictable (if it is possible), for example, it is recommended to mark in the text the next word about which the system will ask.

Natural representation of data is very important, i.e., a user should see a text as if he is using a text editor. We consider that it is difficult for a user to react to suddenly popping up dialogues. At any time, a user should have an opportunity to obtain information about any element of text (words or elements of marking) in an easy way, for example, with a mouse click. Also, there should be an option that allows viewing the marks added to the text or hiding them.

Since a user is restricted to viewing texts on a computer screen, it is important to use multimodal representation. We suggest the usage of different colors or fonts for encoding of different marking elements. If there are too many possible elements of marking, then only the principal ones should be coded with colors. It seems that having more than ten colors with different meaning at the screen makes it too difficult to distinguish between them. At the same time, some too bright colors, like yellow or cyan are not recommended for usage.

The other opportunity for using multimodality is invoking audio, i.e., the system can play or synthesize some audio files pronouncing a user's choice. Thus, a user has an additional chance to note when the audio information contradicts to his decision, i.e., he has made a mistake.

3 Workplace Environment for Spanish

At the present stage of development of the linguist's intelligent workspace for Spanish language, we implemented it for the morphological level. The described environment is based on the morphological analyzer for Spanish AGME [3, 5, 6, 9].

Similar environment for word sense disambiguation is described in [7], though in that system multimodality is not used and the environment is much less user-friendly.

The environment performs morphological marking of texts according to the scheme described above, i.e., the marks are placed automatically when there is no morphological ambiguity, while in case of ambiguity the user should make a selection of the appropriate grammar mark from the list of all possible marks. At the morphological level, marks are parts of speech (noun, verb, etc.) or complete descriptions of grammar categories. This depends on the mode of the system.

The system analyzes text word by word and invokes morphological analysis that determines grammar category/categories and lemma/lemmas[1], for example, for a word *trabajo* (*work*) the result is "*noun-trabajo* (*work*); *verb-trabajar* (*to work*)". In this case there exist homonymy (ambiguity) and the list for selection will contain both variants of analysis.

In Fig. 1 the interface window of the environment is presented.

Environment allows downloading a text from a file and presents it to a user in a natural way. In the left part of the window the menu containing the list of possible grammar categories is displayed. In Fig.1 it can be seen that a word *submarino* (*submarine*) has two possible grammar categories: noun (*sust*) or adjective (*adj*). User can choose the necessary category and click OK button or make a double mouse click on a desired grammar category. The system passes to the next word that has morphological ambiguity. Morphological marks are saved as part of the text, but are not showed to user directly if the corresponding mode is not chosen (as we stated in general considerations above).

Morphologically marked words are encoded by different colors. As can be seen, the processed words have different colors: nouns are green, adjectives are red, verbs are blue, etc. The primary colors are used for most frequent grammar categories: verbs, nouns, and adjectives. Auxiliary words (prepositions, articles, etc.) normally have special treatment because sometimes it is sufficient to know that the word is auxiliary without any further distinctions. Nevertheless, special treatment of auxiliary words is an option and can be switched off.

It is recommended to perform two passes during processing, so, there are two different interface buttons for processing. At the first pass, the words that do not have morphological ambiguity are marked, and only at the second pass words with ambiguity are processed. In this manner, the words that have ambiguity will not be marked after the first pass. Thus, they will be easily distinguishable during processing, so that a user can easily identify the next word that the system will present him at the next moment, as we formulated above in Section 2.

---

[1] Let us remind that lemma is normalized form of a word that usually appears as a headword in the dictionaries.

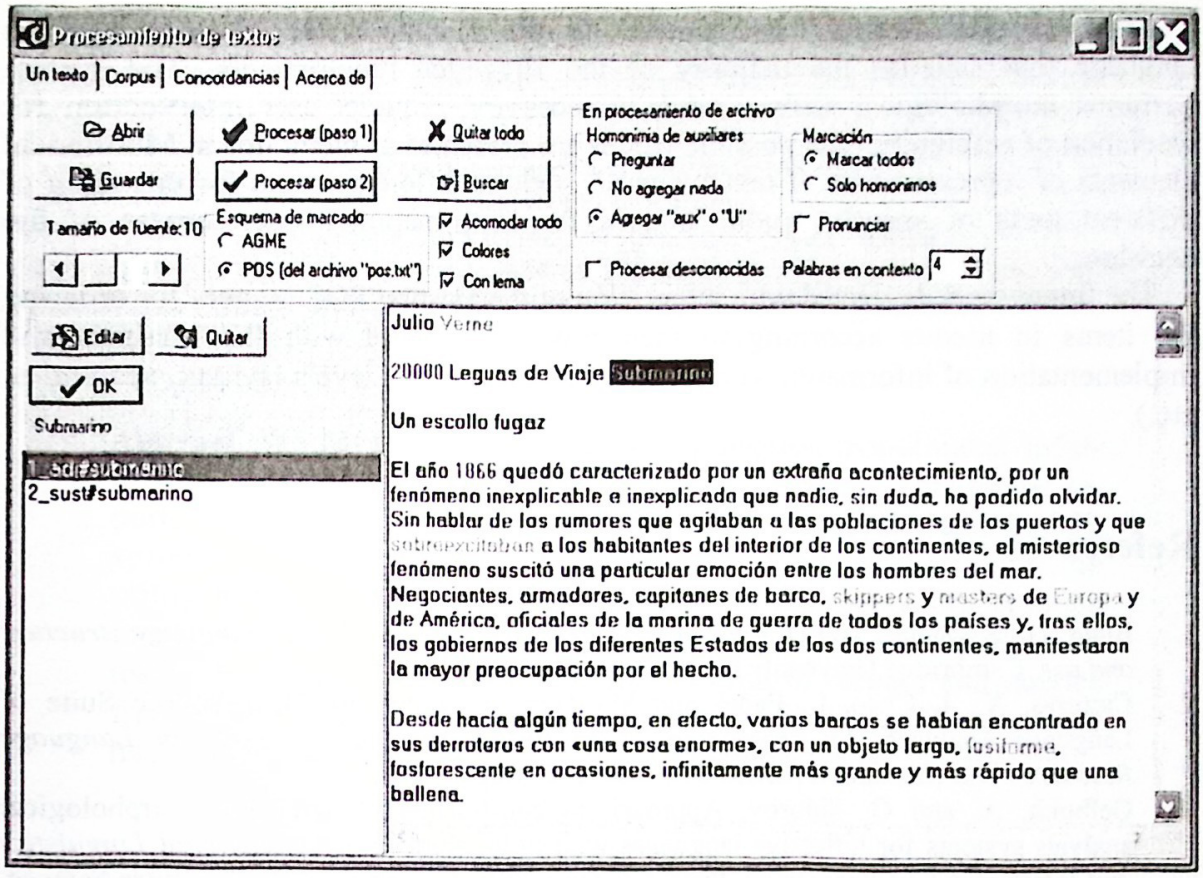

Fig. 1. Interface window (text; menu with the list of categories; options).

System has two modes of encoding of morphological information: only parts of speech (verbs, nouns, etc.) or complete grammar information as detected by the morphological analyzer, see [6]. In the last case, the colors are used only for encoding of the part of speech information, since there are too many possible variants for complete grammar information. Also, there is an option that allows choosing if lemma (normalized form) is added or omitted. In the list in Fig. 1 the lemmas are added after the special symbol #.

4 Conclusions and Future Work

We discussed the requirements that the linguist's intelligent workplace should satisfy:

- Several marking schemes depending on the language level,
- Interface clarity,
- Natural representation of text,
- Multimodality (usage of colors and fonts, audio),
- Integration of additional NLP techniques, and
- Interaction with user on the basis of menus composed of possible language values (marks).

Also, we presented an environment for the morphological level of Spanish language that satisfies the majority of the suggested requirements. The system performs morphological analysis and, if necessary, requests user intervention for resolution of ambiguity. The possible values are presented as menu items. Multimodal elements of representation of text are used, such as different colors for encoding of different parts of speech. Audio is used for confirmation of correctness of the decision.

The future work is related with integration of the existing POS-taggers for ordering the items in menus according to their probabilities and with the analysis and implementation of information encoding for other language levels (syntax, semantics, etc.).

References

1. Biber, D., S. Conrad, and D. Reppen. *Corpus linguistics. Investigating language structure and use.* Cambridge University Press, Cambridge, 1998.
2. Carreras, X., I. Chao, L. Padró and M. Padró. FreeLing: An Open-Source Suite of Language Analyzers. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04).* Lisbon, Portugal. 2004.
3. Gelbukh, A. and G. Sidorov. Approach to construction of automatic morphological analysis systems for inflective languages with little effort. In: *Computational Linguistics and Intelligent Text Processing. Proc. CICLing-2003, Lecture Notes in Computer Science,* N 2588, Springer-Verlag, pp. 215–220.
4. Gelbukh, A., G. Sidorov, SangYong Han, and L. Chanona-Hernandez. Automatic evaluation of quality of an explanatory dictionary by comparison of word senses. *Lecture Notes in Computer Science,* N 2890, 2003, Springer-Verlag, p. 555-561.
5. Gelbukh, A. y G. Sidorov. Analizador morfológico disponible: un recurso importante para PLN en español. En*: Memorias de talleres del congreso internacional Iberamia-2004,* Puebla, México, 2004, pp. 209-216.
6. Gel'bukh, A. Effective implementation of morphology model for an inflectional natural language. *J. Automatic Documentation and Mathematical Linguistics,* Allerton Press, vol. 26, N 1, 1992, pp. 22–31.
7. Ledo Mezquita, Y., G. Sidorov, and A. Gelbukh. Tool for Computer-Aided Spanish Word Sense Disambiguation. *Lecture Notes in Computer Science,* N 2588, 2003, Springer-Verlag, pp. 277–280.
8. Pazos R., Rodolfo A., A. Gelbukh, J. Javier González, E. Alarcón, A. Mendoza, P. Domínguez. Spanish Natural Language Interface for a Relational Database Querying System. *Lecture Notes in Artificial Intelligence,* N 2448, 2002, Springer-Verlag, pp. 123-130.
9. Sidorov, G. O. Lemmatization in the automatized system for compilation of personal style dictionaries of literary writers. In: *Word by Dostoyevsky.* Moscow, Russia, Russian Academy of Sciences, 1996, pp. 266-300.