A Proposal for a Spanish Surface Realization Shared Task

Pablo Gervás and Miguel Ballesteros
Departamento de Ingeniería
del Software e Inteligencia Artificial
Universidad Complutense de Madrid
Madrid, 28040 Spain
pgervas@sip.ucm.es, miballes@fdi.ucm.es

Abstract

We propose a competitive shared evaluation task for Surface Realization in Spanish. The task would be carried out in 2012. It would involve the generation of text in Spanish from a common ground input shared by all systems. Separate corpora for training/development (composed of pairs of common ground input and expected string result) and testing (only common ground input) will be provided. Automatic evaluation procedures will be provided. Submitted results will also be subject to human evaluation. The present proposal is tentative in two different ways. First, the authors intend to revise the proposal in view of the experience and feedback of the Surface Realization Pilot Task currently in process for English, once its results are made public (due in September, 2011). Second, the authors are willing to colaborate both with organizers of equivalent tasks for other languages or more researchers interested in surface realization for Spanish.

1 Background

Two main arguments motivate this proposal: the generally accepted need of establishing comparative forms of evaluation for NLG, and the overarching trend in NLP to extend tools and resources to languages other than English. In the context of the present call for proposals it should not be necessary to argue in favour of the first motivation. Interested readers can be referred to (Belz et al., 2010) and to the call for proposals itself for cogent argumentation on this point.

Regarding the second motivation, it can be defended by analogy with observed trends in Natural Language Analysis (NLA). The field of NLA has experienced a significant boom as a result of the success of statistical approaches. Crucial to this effort was the development of annotated corpora susceptible of being used for training. The existence of these corpora has made it possible to develop a large number of applications based on machine-learning. But this effort has been restricted to the languages for which corpora were available. This restriction has led to imbalances in the coverage that these tools provide across different languages, with a proliferation of tools for English and scarcity for many other languages. A large number of research efforts have been devoted in recent times to correct this imbalance, with researchers, universities, governments and international institutions focused on extending coverage to other languages. It would be a pity if a similar situation is allowed to arise in the field of NLG. The present call for proposals is designed to contribute to the development of a consensus in the use of comparative forms of evaluation in NLG. As such, it should include as soon as possible an effort to address the issue of extension of these methods to other languages.

Another argument in favour of extending these exploratory efforts to other languages can be found in the proliferation of statistical approaches to Surface Realization. Statistical approaches, in contrast to more knowledge-based approaches, should allow rapid development of solutions for alternative languages with little effort, provided the necessary training corpora are available.

Finally, once comparative forms of evaluation start being available, generic observations concerning relative the merit of different approaches are likely to arise, such as for instance, empirical observations on whether statistical or rule-based approaches perform better. For the sake of completeness, it becomes important that any such observations be well founded on comparative studies across different languages. As an example, constituent-based parsing was prevalent for many years in computational approaches to languages while English was the primary object, and yet lost ground very quickly to dependency-based analyses once languages with more complex word order started to be considered. An effort should be made to avoid any similar oversight in Surface Realization.

Spanish is a good candidate as an alternative language for several reasons. First, it is widely used in the world, so any development efforts are likely to have potential application and a large market. Second, it differs from English sufficiently enough to provide a comparative view point. Issues that may introduce difficulties from the surface realization point of view include: long-range agreement in gender, more complex morphology of verbs, pronouns, nouns and adjectives. Gender agreement is particularly significant in languages were words have lexical gender as well as conceptual gender. In Spanish, different synomyms that refer to the same concept can be masculine or femenine, irrespective of the gender of the concept (which may even be neuter in gender). Spanish requires gender agreement between nouns and any accompanying adjectives, and between subjects and attributes in copulative sentences. Third, resources exist in Spanish that can be used as source for the development of a surface realization corpus. Finally, surface realizers have been developed for Spanish in the past, so it should be possible to compare modern approaches with earlier knowledge-based ones.

On the availability of surface realizers for Spanish, two classic surface realizers – FUF (Elhadad and Robin, 1992) and KPML (Bateman, 1995) – have a version for Spanish. These realizers have been deployed in applied contexts. A version of Surge adapted for Spanish was used for story narration (Callaway et al., 1999) although the coverage is inferior to the English original version. A Spanish version of KPML was applied in a chemistry querying system (Aguado et al., 1998). Melero (2006) combined rule–based approaches and machine–learning approaches for Spanish syntactic generation. This system was developed for a commercial machine–translation, the input was a deep syntactic representation and the output was grammatically aceptable text written in Spanish language.

Section 2 outline specifically our proposal considering the organization, data, evaluation and input/output representations. Finally, Section 3 presents some conclusions.

2 Our Shared Task Proposal

Surface realization normally requires an important quantity of knowledge about the structure of the target language, usually represented as a set of grammar rules or other linguistic constraints. Taking all of this into account and in order to continue providing a common forum for these activities, we propose the tools to include Spanish in a Surface Realization Task. In this Section we discuss specifically our proposal considering the organization, data, evaluation and input/output representations.

2.1 Data to be used

The data sets will be based on the Spanish AnCora corpus that was provided as training set for the CoNLL–2009 shared task. We will process this corpus to obtain a format suitable for the surface realization task.

2.1.1 AnCora Corpus: CoNLL 2009 Shared Task Data

AnCora (Palomar et al., 2004; Taulé et al., 2008) is a multilevel annotated corpus of Spanish texts. It has 528,440 lexical tokens. It is mainly based on newspaper texts with their dependency syntactic annotations, named–entity boundaries and semantic dependencies in Spanish. AnCora was developed by the Clic group at the University of Barcelona and it is annotated with morphological (PoS), syntactic (constituents and functions) and semantic (argument structure and thematic roles, semantic class, named entities and WordNet senses) information. The annotation was performed manually, semiautomatically, or fully automatically, depending on the encoded linguistic information, and it uses as a

source the Cast3lb constituency treebank (Civit et al., 2006). It is the Spanish corpus provided for the CoNLL–2009 shared task[1] on "Syntactic and Semantic Dependencies in Multiple Languages" (Hajič et al., 2009).

We have contacted the Clic research group and we have their approval for carrying out the present proposal. They have suggested we may be able to use a forthcoming revised version of the AnCora corpus.

2.1.2 Our Future Data

Our future SR Task data will be derived from the CoNLL 09 AnCora corpus. We will process and adapt the treebank to make it useful for the generation task. It is expected that the actual format taken by this data wil depend largely on the insights obtained from the Surface Realization Pilot Task for English currently taking place during 2011.

It is worth to emphasize that AnCora contains a wide range of sentence lengths, though most of them are between 20 and 50 wordforms. This provides a good benchmark for a surface realization task, with realizations over a broad range of lengths. Moreover, as it is shown in (Gardent and Kow, 2006) surface realization is exponential in the length of the input. This makes the AnCora corpus very suitable for this proposal. Figure 1 shows the distribution of sentences in the AnCora corpus according to their length.
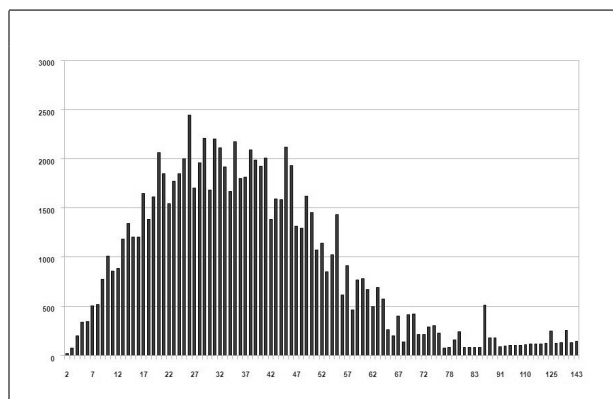

Figure 1: Distribution of sentences in the AnCora corpus according to their length. The x axis represents length and the y axis the approximate number of sentences.

We hope to produce two types of input representations, following the guidelines presented in (Belz et al., 2010), one shallow and one deep. For both shallow and deep representations relations will be randomly sorted and sentences will have single sentence roots.

- **Shallow Dependency Input**

  The shallow representation will include the dependency syntactic tree for every piece of text that is included in the CoNLL'09 data format. The information at each node will consist of a word's lemma, a number and a tense feature, and a coarse–grained POS-tag derived from the AnCora annotation. The edges between nodes will be labeled with the syntactic dependency annotations in the AnCora corpus.

  We have manually developed a transformation to the CoNLL'09 data format into the shallow representation, following the guidelines of the Surface Realisation Shared Task currently taking place in 2011. Figure 2 shows the output of our transformation for the sentence example: *Y, en la mesa, se acabó eso de usar los palillos una sola vez y tirarlos [And, at the table, no more using the toothpicks once and throw them out]*, the representation follows the same structure as the release of the current English Shared Task. It contains only lemmas and shallow dependency relations between nodes.

  If the forthcoming version of the AnCora corpus has a different data format and we decide to use it, we would have to modify the transformation or adapt the data to the expected output.

- **Deep Semantic Input** The deeper representation will be constructed by adding to the Shallow representation the semantic annotation included in the CoNLL'09 data format. Therefore, the information at each node will consist of a word's lemma, a number and a tense feature, and the sense tag (semantic tag). An, as done in (Belz et al., 2010), there will be no POS–tag information. The edges between nodes will be labeled with semantic labels derived from the AnCora annotation for the CoNLL'09 Shared Task.

  For the development of the deep representation, we have contacted Simon Mille and Leo Wanner who are trying to refine AnCora's tagset at

---

```
sentId=1
SROOT    1      0       acabar  CPOS=main        num=s|person=3|tense=past
         et     2      1       y        CPOS=coord
         cc     3      1       en       CPOS=prep
         f      4      3                CPOS=comma
         sn     5      3       mesa     CPOS=common      gen=f|num=s
                spec   6      5       el        CPOS=det         gen=f|num=s
         f      6      3                CPOS=comma
         mprop  7      1       él       CPOS=pron_art    person=3
         suj    8      1       ese      CPOS=pron_dem    num=s
                sp     9      8       de        CPOS=prep
                S      10     9       usar      CPOS=main        tense=infinitive
                cd     11     10      palillo CPOS=common      gen=m|num=p
                        spec   12     11      el        CPOS=det_art     gen=m|num=p
                cc     13     10      vez       CPOS=common      gen=f|num=s
                        spec   14     13      uno       CPOS=det_indef gen=f|num=s
                        s.a    15     13      solo      CPOS=qualif      gen=f|num=s
                coord  16     10      y        CPOS=coord
                S      17     10      tirar     CPOS=main        tense=infinitive
         f      18     1       .        CPOS=punct

Y, en la mesa, se acabó eso de usar los palillos una sola vez y tirarlos.
```

Figure 2: Shallow transformation of the following AnCora sentence: *Y, en la mesa, se acabó eso de usar los palillos una sola vez y tirarlos [And, at the table, no more using the toothpicks once and throw them out]*

the syntactic level (around 60 syntactic tags), and introduce temporary semantic tags in order to facilitate the mapping to the deeper levels (shallow and deep semantics) (Mille and Wanner, 2010). In this way, with their work and the forthcoming version of AnCora (Mariona's work) we should have a robust corpus that will be suitable for the generation task.

2.2 Evaluation

Evaluating surface realization is intrinsically difficult, due to the fact that there is usually no a single correct answer, but rather a range of possible correct answers, some of them better than others. To address this problem, based on the data resources described above we intend to develop evaluation techniques based on Fluency, Clarity and Appropriateness that take this difficulty into account. To this end, outputs will be evaluated by a variety of automatic metrics and human–assessed quality criteria.

We intend to revise this aspect of the proposal based on feedback from the SR pilot task for English. In principle, the evaluation techniques and methodology developed for English should be applicable to Spanish with little or no modification.

3 Conclusion

In this paper, we have proposed a Shared Task for Surface Realization of Spanish, as done in (Belz et

al., 2010). The aim of the proposal is to extend the resources and techniques developed this year for the English Surface Realization Shared Task 2011 to a different language. This should test the techniques beyond the scope for which they were developed and provide resources for the development of surface realizers in a different language.

This proposal could be undertaken as a stand alone task, in tandem with the second iteration of the surface realization task for English, or as part of a multilingual shared task for surface realization. In general terms, the spirit of this proposal is that the use of languages, other than English, in NLG should be promoted. The authors are willing and qualified to provide or recruit the knowledge necessary to build the Spanish data and to evaluate the different participant systems.

We have developed a webpage[2], in which we explain our proposal and we invite people to collaborate in the task. In response to a recent call for expression of interest in this task we have received replies from research groups interested both in submitting and in colaborating in the development of the task.

Acknowledgments

_____

[2]http://nil.fdi.ucm.es/srspa

References

G. Aguado, J. Bateman, S. Bernardos, M. Fernández, A. Gómez-Pérez, E. Nieto, A. Olalla, and A. Sánchez. 1998. Ontogeneration: Reusing domain and linguistic ontologies for spanish text generation. In *Proc. ECAI. Workshop on Applications of Ontologies and Problem Solving Methods*.

John A. Bateman. 1995. Basic technology for multilingual theory and practise: the kpml development enviroment. In *Proc. IJCAI-95 Workshop on MULTILINGUAL TEXT GENERATION*, pages 1–12.

Anja Belz, Mike White, Josef van Genabith, Deirdre Hogan, and Amanda Stent. 2010. Finding common ground: towards a surface realisation shared task. In *Proceedings of the 6th International Natural Language Generation Conference*, INLG '10, pages 268–272, Stroudsburg, PA, USA. Association for Computational Linguistics.

Charles B. Callaway, Brent H. Daniel, and James C. Lester. 1999. Multilingual natural language generation for 3d learning environments. In *In Proceedings of the 1999 Argentine Symposium on Artificial Intelligence, pages 177190, Buenos Aires*, pages 177–190.

Montserrat Civit, Maria Antònia Martí, and Núria Bufí. 2006. Cat3lb and cast3lb: From constituents to dependencies. In *FinTAL*, pages 141–152.

Michael Elhadad and Jacques Robin. 1992. Controlling content realization with functional unification grammars. In *Proceedings of the 6th International Workshop on Natural Language Generation: Aspects of Automated Natural Language Generation*, pages 89–104, London, UK. Springer-Verlag.

Claire Gardent and Eric Kow. 2006. Three reasons to adopt tag-based surface realisation. In *Proceedings of the Eighth International Workshop on Tree Adjoining Grammar and Related Formalisms*, TAGRF '06, pages 97–102, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009), June 4-5*, Boulder, Colorado, USA.

Maria Teresa Melero Nogués. 2006. *Combining machine learning and rule-based approaches in Spanish syntactic generation*. Ph.D. thesis.

Simon Mille and Leo Wanner. 2010. Syntactic dependencies for multilingual and multilevel corpus annotation. Valletta (Malta), 05/2010.

M. Palomar, Montserrat Civit, A. Díaz, L. Moreno, E. Bisbal, M. Aranzabe, A. Ageno, M.Antonia Martí, and Borja Navarro. 2004. 3lb: Construcción de una base de datos de árboles sintáctico–semánticos para el catalán, euskera y español. In *Proceedings of the XX Conference of the Spanish Society for Natural Language Processing (SEPLN)*, pages 81–88. Sociedad Española para el Procesamiento del Lenguaje Natural.

Mariona Taulé, M.Antonia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of 6th International Conference on Language Resources and Evaluation*.