Adapting SimpleNLG to Spanish

Alejandro Ramos-Soto and **Julio Janeiro-Gallardo** and **Alberto Bugarín**
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
University of Santiago de Compostela, Spain
{alejandro.ramos,julio.janeiro,alberto.bugarin.diz}@usc.es

Abstract

We describe SimpleNLG-ES, an adaptation of the SimpleNLG realization library for the Spanish language. Our implementation is based on the bilingual English-French SimpleNLG-EnFr adaptation. The library has been tested using a battery of examples that ensure that the most common syntax, morphology and orthography rules for Spanish are met. The library is currently being used in three different projects for the development of data-to-text systems in the meteorological, statistical data information, and business intelligence application domains.

1 Introduction

In recent times, natural language generation (NLG) is receiving increased attention beyond its research community. Commercial success is nowadays a fact for several NLG companies and this trend is likely to expand in coming years, as many opportunities will arise for the development of NLG systems that provide useful interpretable information and address different needs within organizations.

In this context, software for NLG purposes which is freely available is an exception rather than a norm. In fact, according to Reiter (Reiter, 2017), this problem is worsened by a lack of visibility of available software. In other words, there is not much software for NLG and most of it is unknown to the wide public. However, among the different tasks that can be identified within an NLG system according to (Reiter and Dale, 2000), software for the realization task (the actual text generation process from the intermediate representation of the information to be conveyed) is actually plentiful when compared to other tasks like referring expression generation.

In this paper we focus on a specific realizer for the English language, namely SimpleNLG, which was originally presented in (Gatt and Reiter, 2009), but has been improved over the years and has become a popular and active project. As a Java library, SimpleNLG honours its name by providing an intuitive object-oriented API for generating text from a lexical-syntactic specification.

Moreover, different adaptations of SimpleNLG have been made to support other languages. Among them, the bilingual English-French version developed by Vaudry (Vaudry and Lapalme, 2013), provided an abstraction of the common linguistic features shared by both languages. Subsequent adaptations include German (Bollmann, 2011), Brazilian Portuguese (de Oliveira and Sripada, 2014) and Italian (Mazzei et al., 2016), based on different versions of the original and the bilingual versions of the library. SimpleNLG has also served as inspiration for a realizer for the Indian Telugu language (Dokkara et al., 2015).

2 Adapting SimpleNLG-EnFr to Spanish

Given that Spanish is one of the most spoken languages in the world in terms of native speakers, there are many potential benefits about providing a realizer that supports this language. To our knowledge, there has been a previous attempt at developing a Spanish version of SimpleNLG, as in (Trzpis, 2015). However, this project was not made publicly available and, according to its description, several fea-

tures were missing or incomplete, such as interrogative questions, gerund and participle verb forms or an extensive lexicon.

In order to address this, we provide an adaptation of the bilingual English-French version of SimpleNLG for the Spanish language. Specifically, we have opted for the creation of a bilingual English-Spanish version that follows the original class structure of SimpleNLG-EnFr, instead of adding a third language to the existing implementation of SimpleNLG-EnFr. While we believe that creating a more general framework based on SimpleNLG with multilingual support is a necessary task that should be undertaken in the future, our main interest is to provide an useful NLG tool which is not currently available for NLG developers that may need use the Spanish language.

Thus, as in SimpleNLG-EnFr, most of the basic framework is shared between both languages, such as document elements and some grammar rules which are common for both English and Spanish. Based on the abstract classes, we have subclassed several functionalities regarding syntax, morphology, and ortography and adapted an already existing lexicon which provides a very extensive collection of Spanish words and associated inflections. The Spanish grammar used as reference is the *"Nueva gramática de la lengua española"* (RAE, 2011).

2.1 Lexicon

Instead of building our own lexicon, we opted for reusing already well-known and reputed existing resources in the literature. In particular, we have used the Spanish dictionary provided by the FreeLing project (Padró and Stanilovsky, 2012) to develop and test SimpleNLG-ES. This dictionary provides 555,000 forms corresponding to more than 76.000 lemma-PoS (Part of Speech) combinations. Since the dictionary cannot be used directly with SimpleNLG, we converted the original file format into an XML file format which is compatible with the realizer.

3 Features of the Spanish language

We describe here some of the most interesting features of the Spanish language which have been incorporated into SimpleNLG-ES. These involve syntax, orthography, morphology and morphophonology elements.

3.1 Syntax

Some of the most relevant features in terms of syntax include the adjective positioning, the use of reflexive pronouns for passive clauses, relatively simple interrogative clauses, and the subjunctive conversion of imperative statements that are used as relative clauses.

Verb phrase: Verb phrases are in structure similar to English verb phrases. Among the most interesting verb phrase features of the Spanish language for which we have added support, the construction of verb clauses without subject using the verb *"haber"*, is expressed as *"hay"* (which actually means *"there is/are"*). Likewise, passive clauses, which are less common in Spanish, can be constructed by adding the reflexive pronoun *"se"*. For instance, *"se crearon tres prototipos"* means *"three prototypes were created"*.

Noun phrase: Noun phrases follow a similar pattern to their English counterpart, with a determiner, a noun, and one or more adjectives. The main difference is that adjectives are usually positioned after the noun, although they can also be positioned before the noun for emphatic purposes or slight differences in meaning. For instance, *"that **good man**"* can be expressed as *"ese hombre **bueno**"* or *"ese **buen** hombre"*.

A specific feature involving noun phrases when used as indirect objects is that a preposition *'a'* has to be attached before the noun phrase. For instance, *"I gave **the kids** a toy"* would be expressed as *"Le di un juguete **a los niños**"*. This actually coincides with the less used English alternative, *"I gave a toy **to the kids**"*.

Interrogative clause: Interrogative clauses are formed in many cases by simply adding the proper punctuation signs at the beginning and end of a sentence. For instance, a simple statement converted into a yes/no question like *"you want to eat"* → *"do you want to eat?"* would be expressed as *"quieres comer"* → *"¿quieres comer?"*.

In the case of a 'who' question, where this particle plays the role of an indirect object, the beginning of

the question incorporates a preposition 'a', just like in the indirect object case described above for the noun phrase. For instance, *"**Who** do you prefer?"* would be expressed as *"¿A **quién** prefieres?"*.

Relative clause: In the case of relative clauses, subordinate imperative sentences are transformed into subjunctive. For instance, *"Throw away the stones."* → *"I want you to throw away the stones."*, would be expressed as *"Tira las piedras."* → *"Quiero que **tires** las piedras."* (*"I want that you throw away the stones"*, literally translated).

3.2 Orthography

Most general orthography rules that are used in English also apply to Spanish, such as beginning sentences with capital letters or punctuation using commas and dots. Thus, in SimpleNLG-ES most orthography code is shared for both languages.

In Spanish there exist orthographic rules that determine whether specific letters within words must include a tilde, a diacritical mark that helps determine the word pronunciation. For example, in 'pretérito' the tilde over the 'e' letter implies a stress in the pronunciation for that syllable. In our case, the Freeling lexicon already provides the correct forms of the words, including tildes. This allowed us to avoid including the related orthography rules for determining tildes, which involve identifying syllables and ending letters for words.

In fact, the only Spanish orthographic rule we had to add is the inclusion of the interrogative particle '¿' that marks the beginning of an interrogative sentence.

3.3 Morphology

Gender and number: Determiners, adjectives and some nouns must be inflected in gender and number. In our case the lexicon supports all inflections for base words, but we have also included several rules for inflection of plurals for regular nouns.

Verb tenses: As in other languages, in Spanish there exist regular and irregular verbs. We have included in SimpleNLG-ES the standard rules for inflecting regular verbs, while irregular verbal forms are provided by the lexicon. Moreover, verbs in Spanish can be inflected for several modalities (indicative, subjunctive, imperative) and many different tenses. Tenses involve simple and compound forms using the auxiliar verb *"haber"* (to have). All of them are correctly supported by SimpleNLG-ES.

In order to illustrate the complexity of the verbal system in Spanish, which is supported by SimpleNLG-ES, consider the English sentence *"I **had eaten**"*. In Spanish, depending on the context, this could correspond to the indicative compound tense named *pretérito pluscuamperfecto "yo **había comido**"* or to the *preterito perfecto compuesto "yo **hube comido**"*. At the same time, *"if I **had eaten**"* would correspond to the subjunctive *pretérito pluscuamperfecto "si yo **hubiera comido**"* or *"si yo **hubiese comido**"* (this tense admits two different forms). Thus, depending on the context of the sentence, *"I had eaten"* could correspond in Spanish to three different tenses, four different realizations, and two different modalities.

3.4 Morphophonology

The Spanish language is not as rich as other close languages in terms of morphophonology rules. However, there exist two main contractions between the prepositions *'a'* and *'de'*, and the masculine singular determinant *'el'* that are always used:

- *"a el"* → *'al'* (to the)
- *"de el"* → *'del'* (of the)

Other contractions are used in a more colloquial spoken context, such as *"para atrás"* → *"patrás"* (backwards) and *"para adelante"* → *"palante"* (forwards), but these are not officially recognized by the Real Academia Española (the official organism that regulates the Spanish language). Thus, SimpleNLG-ES includes only support for *'al'* and *'del'*.

4 Test and use

We have tested our adaptation of SimpleNLG for Spanish in different ways. Firstly, we adapted the existing tests for the English language version, as many of the features tested also apply to Spanish. These tests, built using a unit testing framework, check the functioning of the library at different levels, by comparing if the text strings generated by the library match the correct expected results. SimpleNLG-ES passed all these tests successfully.

Secondly, we tested SimpleNLG-ES in the real data-to-text service GALiWeather (Ramos-Soto et al., 2015), which is a template-based NLG system that was deployed in May 2015 as a public service for the Official Meteorology Agency (MeteoGalicia, 2000) of Galicia (NW Spain). GALiWeather produces automatically daily operational weather forecasts for each of the 314 municipalities in Galicia.

Specifically, we refactorized and adapted GALiWeather, to perform realization using SimpleNLG for Spanish. Tests consisted of 76 different real weather forecasts produced since November 2016, which were generated for the same input data. The corresponding forecast texts were generated using both approaches (GALiWeather's original template-based and SimpleNLG-ES realizations) and their strings were matched using unit testing assertion methods.

Only in 7 out of the 76 testing examples (9%), the same non-relevant difference was found between pairs of texts generated for the same input data. This corresponded to the case where the Spanish reflexive pronoun *'se'* appeared. In Spanish, this pronoun can be placed separately or attached to a verb with no changes in meaning, as in *"se podrán encontrar"* or *"podrán encontrarse"*. The templates use the latter case, while our adaptation of SimpleNLG applies the former rule:

- Expected (template-based GALiWeather): *"Se espera que los cielos alternen periodos muy nubosos con otros parcialmente nubosos, aunque ocasionalmente* ***podrán encontrarse*** *poco nubosos o despejados."*
- Actual (SimpleNLG-ES): *"Se espera que los cielos alternen periodos muy nubosos con otros parcialmente nubosos, aunque ocasionalmente* ***se podrán encontrar*** *poco nubosos o despejados."*

In order to further test and refine the library, we are also currently using SimpleNLG-ES in three different projects for the development of data-to-text systems in new domains, with satisfactory results. The first project is in the environmental information domain, also in cooperation with the Galician Meteorology Agency, for automatically generating textual meteorological warnings following the guidelines of the European Meteoalarm service (EUMETNet, 2007). The second project focuses on providing textual descriptions and explanations in natural language about a number of official statistical data and indicators. Finally, the third project is in the business intelligence information realm.

5 Documentation and release

SimpleNLG-ES has been thoroughly documented. Specifically, the source code maintains the original documentation style of SimpleNLG in English. We have also adapted and translated the original documentation and tutorial into Spanish.

Our plan is to release the source code and its associated documentation on GitHub or a similar repository, in order to meet the conditions established by the original Mozilla Public License under which SimpleNLG is published (since its 4.0 version).

6 Conclusions

We have described in this paper the most relevant features of our adaptation SimpleNLG-ES of the realizer SimpleNLG for the Spanish language. This version provides an extensive support for the most common usage of Spanish, using a comprehensive lexicon that covers most of the common Spanish vocabulary and all its inflections.

SimpleNLG-ES has been tested through various means. Moreover, the library is currently being used in three different projects for the development of real D2T systems in the meteorological, statistical data information, and business intelligence application domains. As current and near future work, we will extend the library to also support the Galician language.

Acknowledgments

This work has been funded by TIN2014-56633-C3-1-R and TIN2014-56633-C3-3-R projects from the Spanish "Ministerio de Economía y Competitividad" and by the "Consellería de Cultura, Educación e Ordenación Universitaria" (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF). A. Ramos-Soto is funded by the "Consellería de Cultura, Educación e Ordenación Universitaria" (under the Postdoctoral Fellowship accreditation ED481B 2017/030).

References

Marcel Bollmann, 2011. *Proceedings of the 13th European Workshop on Natural Language Generation*, chapter Adapting SimpleNLG to German, pages 133–138. Association for Computational Linguistics.

Rodrigo de Oliveira and Somayajulu Sripada, 2014. *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, chapter Adapting SimpleNLG for Brazilian Portuguese realisation, pages 93–94. Association for Computational Linguistics.

Sekhar Sasi Raja Dokkara, Verma Suresh Penumathsa, and Gowri Somayajulu Sripada, 2015. *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, chapter A Simple Surface Realization Engine for Telugu, pages 1–8. Association for Computational Linguistics.

EUMETNet. 2007. Meteoalarm website. www.meteoalarm.eu. `http://www.meteoalarm.eu`. Accessed: 2017-05-12.

Albert Gatt and Ehud Reiter, 2009. *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, chapter SimpleNLG: A Realisation Engine for Practical Applications, pages 90–93. Association for Computational Linguistics.

Alessandro Mazzei, Cristina Battaglino, and Cristina Bosco, 2016. *Proceedings of the 9th International Natural Language Generation conference*, chapter SimpleNLG-IT: adapting SimpleNLG to Italian, pages 184–192. Association for Computational Linguistics.

MeteoGalicia. 2000. Meteogalicia operational forecasting website for municipalities. www.meteogalicia.gal. `http://www.meteogalicia.gal/web/predicion/localidades/localidadesIndex.action?idZona=15078`. Accessed: 2017-05-12.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

RAE. 2011. Nueva gramática de la lengua española. `http://www.rae.es/recursos/gramatica/nueva-gramatica`. Accessed: 2017-05-10.

A. Ramos-Soto, A. Bugarín, S. Barro, and J. Taboada. 2015. Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data. *IEEE Transactions on Fuzzy Systems*, 23(1):44 – 57.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Ehud Reiter. 2017. Why isnt there more open-source NLG software? `https://ehudreiter.com/2017/03/17/open-source-nlg-software/`. Accessed: 2017-05-10.

Damian Jozef Trzpis. 2015. Adaptación de una herramienta de generación de lenguaje natural al idioma español. Master Thesis.

Pierre-Luc Vaudry and Guy Lapalme, 2013. *Proceedings of the 14th European Workshop on Natural Language Generation*, chapter Adapting SimpleNLG for Bilingual English-French Realisation, pages 183–187. Association for Computational Linguistics.