Adapting SimpleNLG to Galician language

Andrea Cascallar-Fuentes[1], Alejandro Ramos-Soto[1,2], and Alberto Bugarín[1]

[1]Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS),
Universidade de Santiago de Compostela, Spain
{andrea.cascallar.fuentes,alejandro.ramos,alberto.bugarin.diz}@usc.es
[2]Department of Computing Science, University of Aberdeen
alejandro.soto@abdn.ac.uk

Abstract

In this paper, we describe SimpleNLG-GL, an adaptation of the linguistic realisation SimpleNLG library for the Galician language. This implementation is derived from SimpleNLG-ES, the English-Spanish version of this library. It has been tested using a battery of examples which covers the most common rules for Galician.

1 Introduction

Realisation is the final task in natural language generation. Its goal is to ensure that well-formed texts are generated according to the grammar rules of the output language. Consequently, having tools that facilitate this task is desirable for any developer of a NLG system. For instance, templates are a widely-used realisation mechanism which is appropriate for many application domains, where generated texts are rather static.

Templates, however, are harder to maintain as they grow, and ensuring consistency among the elements of a realised template might become more difficult as more dynamic components appear. To address this kind of issues, other realisation tools pack language rules and syntactic structures to provide a framework for building well-formed sentences. This is the case of SimpleNLG, a Java realiser for English presented in (Gatt and Reiter, 2009) to facilitate realisation tasks. Some versions of this library have been created to support different languages: English-French (Vaudry and Lapalme, 2013), Italian (Mazzei et al., 2016), Brazilian Portuguese (de Oliveira and Sripada, 2014), German (Bollmann, 2011) and English-Spanish (Ramos-Soto et al., 2017). Other realisers described in the literature are Alethgen (Coch, 1996), FUF/SURGE (Elhadad and Robin, 1996), RealPro (Lavoie and Rainbow, 1997), KPML (Bateman, 1997), YAG (McRoy et al., 2000), HALogen (Langkilde-Geary, 2002) and OpenCCG (White, 2006).

This paper describes SimpleNLG-GL, a trilingual realisation tool for English, Spanish and Galician, derived from SimpleNLG-ES (Ramos-Soto et al., 2017). The Galician language is mainly spoken by approximately a million people in Galicia, NW of Spain. It is also closely related to the Portuguese language, since until the Middle Ages both were a single linguistic unit.

Given the closeness of Spanish and Galician, we decided to base this adaptation of SimpleNLG on the dual English-Spanish version. Nevertheless, Galician has a rich variety of specific features that clearly demanded a new adaptation of the library. Thus, we will also show some examples of the necessary steps to translate a phrase from Spanish to Galician, in order to illustrate the higher complexity that the Galician language has with respect to Spanish, and how this influenced our implementation of SimpleNLG-GL.

2 Covered subset of Galician

The Galician grammar used as reference is *"Normas ortográficas e morfolóxicas do idioma galego"* (Galega, 2012), which was created by the *Real Academia Galega* (Royal Galician Language Academy, founded 1906), a scientific institution whose objective is studying the Galician culture and, in particular, its language. This grammar was created to define the orthographic and morphological rules of the Galician language.

2.1 Lexicon

To create the lexicon used to develop this version of SimpleNLG, we chose the Galician dictionary provided by the FreeLing Project (Padró and Stanilovsky, 2012), an open source language

Proceedings of The 11th International Natural Language Generation Conference, pages 67–72,
Tilburg, The Netherlands, November 5-8, 2018.

analysis tool suite which provides some language analysis capabilities for a wide range of languages. This dictionary cannot be used directly by SimpleNLG, so we produced a compatible XML dictionary generated from the original file.

3 Features of the Galician language

In this section, we describe the most interesting features of the Galician language covered by the library, including syntax, orthography and morphology.

3.1 Syntax

3.1.1 Noun phrases

The structure of noun phrases is composed by a determiner, zero or one possessive, a noun and optionally one or more adjectives. When a phrase contains a possessive, in most cases it also includes a determiner (before). For instance, *"o meu fogar"* is translated as *"my home"* when *"o"* means *"the"*, *"meu"* means *"my"* (masculine) and *"fogar"* means *"home"*. Therefore, the literal translation is *"the my home"*. In a phrase with adjectives, the meaning of the phrase can slightly change depending on where the adjectives are placed (before or after the noun). Adjectives after the noun refer to features which were previously unknown by the speaker. However, if the adjective goes before the noun, the referred feature was already known. For instance, *"o novo fogar"* or *"o fogar novo"* mean *"the new home"*. To say *"his new home"* we can express it as *"o seu novo fogar"* or *"o seu fogar novo"*.

A specific feature when noun phrases are used as indirect objects is that a preposition *"a"*, which means *"to"*, is utilised before the phrase and a contraction is generated formed by that preposition and the phrase's determiner if applicable. For instance, *"Eu vin a Victoria"* means *"I saw Victoria"*. An example with contraction is *"Eu vin ao teu gato"*, translated as *"I saw your (male) cat"* when the preposition *"a"* and the masculine determiner *"o"* are contracted forming *"ao"*. Other example, with a feminine noun is *Eu vin á túa gata"* which means *"I saw your cat (female)"*.

3.1.2 Verb phrases

A general structure of verb phrases is composed by a subject, a verb and zero or more objects. However, in Galician there are sentences without a subject using the verb *"haber"* in its third person singular conjugation in the simple tense form *"hai"*, which means *"there is/are"*. For instance *"there is a cat on the tree"* would be expressed as *"hai un gato na árbore"*. Similarly, passive sentences are created adding the reflexive pronoun *"se"* connected to the verb and after it. For instance *"the event was cancelled"* can be expressed as *"cancelou**se** o evento"* (Figure 1).

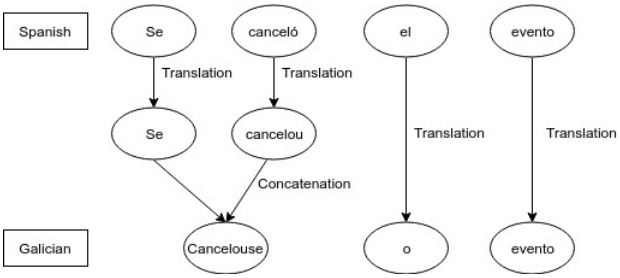

Figure 1: Steps to translate from Spanish to Galician language.

A feature of the Galician language is the pronoun placement in relation to the verb when it is used as a direct or indirect object, either before or after the verb, appearing both combined in the latter case. This collocation depends on the sentence type, e.g., the pronoun is generally placed after the verb in affirmative sentences, whereas it is placed before it in negative sentences. For instance, *"el deume un regalo"* is translated as *"he gave me a present"*. In this case the verb is *"deu"* (*"gave"*) and the pronoun is *"me"*, which is combined with the verb. In a negative sentence, *"el non me deu un regalo"* translated as *"he did not give me a present"*, an the pronoun appears separately before the verb. To handle this feature, the library has to perform the following three tasks:

- Analyse the phrase type. The general rule is that pronouns are placed after the verb, however, we must analyse the phrase to determine its position. Some words change the verb's position as negation adverbs (*"non o vin"* which means *"I did not see it"*), doubt adverbs (*"quizais ela te chame mañá"* translated as *"maybe she calls you tomorrow"*), interrogative pronouns (*"que che pasou?"* which means *"what happened to you?"*).

- Split the verb into syllables. Adding the pronoun to the verb, its accentuation can change and an accent mark has to be added or moved if the verb has it. Therefore, we need to analyse the verb to find out its category according to where its strong syllable is. For instance,

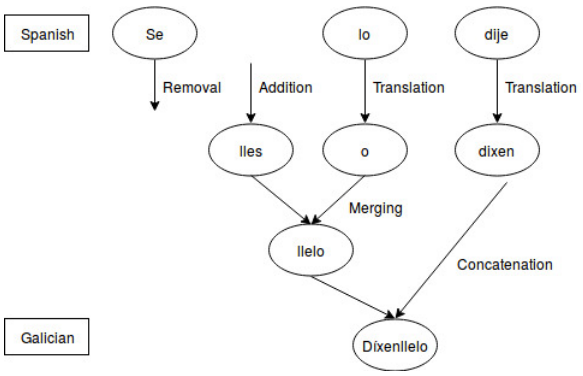

Figure 2: Steps to translate from Spanish to Galician language.

the form *"entendo"*, the first person singular conjugation in the present simple tense form of the verb *"understand"*, is split as *"en-tendo"*.

- Accentuation. Once we know the verb's category, the last step is to concatenate the pronoun to the verb and to check the new word's accentuation, adding, moving or removing an accent mark. For instance, the first person singular conjugation in the past simple tense form *"dixen"* of the verb *"dicir"* (*"say"*) has the stress on its first syllable. If we add the pronoun *"lle"* expressed as *"to him"*, *"to her"* or *"to it"*, the composed word is *'díxenlle'*. If we add the contraction of pronouns (Table 6) *"llelo"*, composed by the pronoun *"lles"*, which means *"to them"*, and the pronoun *"o"*, which means *"it"*, the composed word is *"díxenllelo"* (Figure 2). The stress of the new words in these examples is also their first syllable. However, due to the Galician orthography rules, an accent mark (which is not present in the original word) has to be put on these first syllables.

More details about pronoun concatenation are given in Section 3.3.3.

3.1.3 Interrogative phrases

Interrogative phrases can be formed in many ways by simply adding the punctuation mark at the end of the sentence. For example, *"tes frío"* that means *"you are cold"* can be transformed into a question simply adding the punctuation mark *"tes frío?"* that means *"are you cold?"*.

When the interrogative pronouns *"what"* and *"who"* have the role of indirect objects, the preposition *"a"* is inserted in the beginning of the question. For instance, *"a quen chamaches?"* can be expressed as *"who did you call?"*.

3.2 Orthography

General Galician orthography rules (e.g. punctuation, capital letters at the beginning of sentences...) are like English and Spanish ones. This means that SimpleNLG already has them implemented.

As we mentioned before, Galician has special rules for word categorisation regarding their stress syllables. Besides, some words have accent marks on one vowel to stress the strong syllable. The entries in the lexicon we use contain accents according to the Galician orthography rules. However, it does not contain generated words formed by contractions, so we implemented the corresponding accentuation rules to handle these cases.

3.3 Morphology

3.3.1 Gender and number

Determiners, nouns and adjectives have to be inflected in gender and number. Our lexicon provides the base form of a word but not its gender and number variations so we had to implement some rules to generate them when they are regular.

3.3.2 Verb tenses

Verbs can be regular or irregular. We implemented some rules to generate regular forms, whereas the irregular ones are provided by the lexicon.

3.3.3 Morphophonology

Galician is a very rich language in terms of its morphophonology rules. SimpleNLG-GL implements the contractions that exist between prepositions and articles, and also between pronouns when they function as direct and indirect objects.

Prepositions and articles: The prepositions shown in Table 1 can contract with definite articles, *"o"*, *"a"*, *"os"*, *"as"* which mean *"the"*, for instance *"o gato"* means *"the cat"*; and also indefinite articles, *"un"*, *"unha"*, *"uns"*, *"unhas"* which mean *"a"*, for instance, *"un gato"* means *"a cat"*. These contractions have the following meanings:

- *"a + definite article"* means *"to the"* whereas *"a + indefinite article"* means *"to a"*

Galician	English
a	to
con	with
de	of
en	in
por	by
tras	after

Table 1: Meaning of prepositions

		Articles			
		o	**a**	**os**	**as**
Prepositions	**a**	ao	á	aos	ás
	con	co	ca	cos	cas
	de	do	da	dos	das
	en	no	na	nos	nas
	por	polo	pola	polos	polas
	tras	tralo	trala	tralos	tralas

Table 2: Contractions between prepositions and definite articles

- *"con + definite article"* means *"with the"* whereas *"con + indefinite article"* means *"with a"*
- *"de + definite article"* means *"of the"* whereas *"de + indefinite article"* means *"of a"*
- *"en + definite article"* means *"in the"* whereas *"en + indefinite article"* means *"in a"*
- *"por + definite article"* can mean *"by the"* whereas *"por + indefinite article"* can mean *"by a"*
- *"tras + definite article"* can mean *"after the"* whereas *"tras + indefinite article"* can *"after a"*

In Tables 2 and 3 the contractions are shown.

Pronouns: The atonic pronouns having the role of indirect objects shown in Table 4 can also contract with others which have the role of direct objects which are shown in Table 5. For instance, the phrase *"El deumo"* (Figure 3) is expressed as *"He gave it to me"*, where *"deu"* means *"gave"* and

		Articles			
		un	**unha**	**uns**	**unhas**
Prepositions	**con**	cun	cunha	cuns	cunhas
	de	dun	dunha	duns	dunhas
	en	nun	nunha	nuns	nunhas

Table 3: Contractions between prepositions and indefinite articles

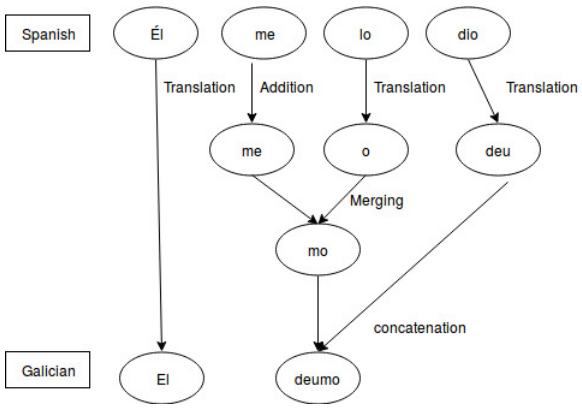

Figure 3: Steps to translate from Spanish to Galician language.

Galician	English
me	to me
che	to you (singular)
lle	to him/her/it
nos	to us
vos	to you (plural)
lles	to them

Table 4: Meaning of atonic pronouns with the role of indirect objects

"mo" is the contraction of *"me"* (*"to me"*) and *"o"* (*"it"*), respectively. In Table 6 all possible combinations are shown.

4 Availability, test and documentation

SimpleNLG-GL is available and fully downloadable at (Cascallar-Fuentes et al., 2018)

The documentation from SimpleNLG-ES has also been adapted to this version, and is also available at the library repository in the form of a wiki, as usual in SimpleNLG, which contains a tutorial with some examples.

SimpleNLG-GL has been tested using 180 unit tests adapted from SimpleNLG-ES. New tests have been generated to cover Galician language features not present in the Spanish language. We

Galician	English
o	him/it
a	her/it
os	them (masculine)
as	them (feminine)

Table 5: Meaning of atonic pronouns with the role of direct objects

	o	a	os	as
me	mo	ma	mos	mas
che	cho	cha	chos	chas
lle	llo	lla	llos	llas
nos	nolo	nola	nolos	nolas
vos	volo	vola	volos	volas
lles	llelo	llela	llelos	llelas

Table 6: Contractions between atonic pronouns

had to create new tests to cover the Galician language features previously described. For instance, we created 11 tests to cover contractions between all prepositions and articles, 4 tests to cover atonic pronouns collocation and 7 tests to cover contractions between atonic pronouns. Also, in some of the adapted tests from the Spanish version these features are present as well.

Besides, SimpleNLG-GL has been used in the real data-to-text service GALiWeather (Ramos-Soto et al., 2015), which generates daily weather forecasts for the Galician municipalities in the Website of the Galician Meteorological Agency (Agency). This service combines a template-based approach with the use of SimpleNLG-GL to generate correct sentences, in terms of the agreement between the elements of the phrase (for example, for choosing correct verb conjugations and ensuring gender and number coherence, among others).

5 Conclusions

We have described SimpleNLG-GL, an adaptation of the SimpleNLG Java realisation engine for Galician language, that provides a sophisticated covering of even the most complex rules found in this language. This library has been tested extensively using unit tests, 180 adapted from SimpleNLG-ES testing, whilst other 22 were newly developed for SimpleNLG-GL.

Acknowledgements

This research was supported by the Spanish Ministry of Economy and Competitiveness (grants TIN2014-56633-C3-1-R and TIN2017-84796-C2-1-R) and the Galician Ministry of Education (grants GRC2014/030 and "accreditation 2016-2019, ED431G/08"). All grants were co-funded by the European Regional Development Fund (ERDF/FEDER program). A. Ramos-Soto is funded by the "Consellería de Cultura, Educación e Ordenación Universitaria" (under the Postdoctoral Fellowship accreditation ED481B 2017/030).

References

Galician Meteorological Agency. Meteogalicia website. http://www.meteogalicia.gal. Accessed: 2018-09-26.

John A. Bateman. 1997. Enabling technology for multilingual natural language generation: the KPML development environment. *Natural Language Engineering*, 3(1):15–55.

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

Andrea Cascallar-Fuentes, Alejandro Ramos-Soto, and Alberto Bugarín. 2018. SimpleNLG-GL on CiTIUS GitHub. https://github.com/citiususc/SimpleNLG-GL. Accessed: 2018-09-26.

José Coch. 1996. Overview of AlethGen. In *Eighth International Natural Language Generation Workshop (Posters and Demonstrations)*, pages 25–28. Association for Computational Linguistics.

Michael Elhadad and Jacques Robin. 1996. An Overview of SURGE: a Reusable Comprehensive Syntactic Realization Component. In *Eighth International Natural Language Generation Workshop (Posters and Demonstrations)*, pages 1–4. Association for Computational Linguistics.

Real Academia Galega. 2012. Normas ortográficas e morfolóxicas do idioma galego. https://academia.gal/documents/10157/704901/Normas+ortogr%C3%A1ficas+e+morfol%C3%B3xicas+do+idioma+galego.pdf. Accessed: 2018-09-26.

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

Irene Langkilde-Geary. 2002. An Empirical Verification of Coverage and Correctness for a General-Purpose Sentence Generator. In *Proceedings of the International Natural Language Generation Conference*, pages 17–24.

Benoit Lavoie and Owen Rainbow. 1997. A Fast and Portable Realizer for Text Generation Systems. In *Proceedings of the 15th conference on Applied natural language processing*, pages 265–268. Association for Computational Linguistics.

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

Susan Weber McRoy, Songsak Channarukul, and Syed S. Ali. 2000. YAG: A Template-Based Generator for Real-Time Systems. In *Proceedings of the 1st international conference on Natural language generation*, pages 264–267. Association for Computational Linguistics.

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

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

A. Ramos-Soto, J. Janeiro-Gallardo, and Alberto Bugarín. 2017. Adapting SimpleNLG to Spanish. In *10th International Conference on Natural Language Generation*, pages 144–148. Association for Computational Linguistics.

Alejandro Ramos-Soto, Alberto José Bugarín Diz, Senén Barro, and Juan Taboada. 2015. Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data. *IEEE Trans. Fuzzy Systems*, 23(1):44–57.

Pierre-Luc Vaudry and Guy Lapalme. 2013. Adapting simplenlg for bilingual english-french realisation. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 183–187, Sofia, Bulgaria. Association for Computational Linguistics.

Michael White. 2006. CCG Chart Realization from Disjunctive Inputs. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 12–19. Association for Computational Linguistics.