

Accepted Manuscript

Please cite this proceedings as: Silvia García-Méndez, Milagros Fernández-Gavilanes, Enrique Costa-Montenegro, Jonathan Juncal-Martínez, and Francisco Javier González-Castaño. 2018. Automatic Natural Language Generation Applied to Alternative and Augmentative Communication for Online Video Content Services using SimpleNLG for Spanish. *In Proceedings of the 15th International Web for All Conference (W4A '18)*. Association for Computing Machinery, New York, NY, USA, Article 19, 1–4. <https://doi.org/10.1145/3192714.3192837>

Link to published version: <https://doi.org/10.1145/3192714.3192837>

General rights:

© 2018 Association for Computing Machinery

Automatic Natural Language Generation Applied to Alternative and Augmentative Communication for Online Video Content Services using *SimpleNLG* for Spanish

Silvia García-Méndez
GTI Research Group,atlanTTic
Vigo, Spain
sgarcia@gti.uvigo.es

Milagros Fernández-Gavilanes
GTI Research Group,atlanTTic
Vigo, Spain
mfgavilanes@gti.uvigo.es

Enrique Costa-Montenegro
GTI Research Group,atlanTTic
Vigo, Spain
kike@gti.uvigo.es

Jonathan Juncal-Martínez
GTI Research Group,atlanTTic
Vigo, Spain
jonijm@gti.uvigo.es

Francisco Javier González-Castaño
GTI Research Group,atlanTTic
Vigo, Spain
javier@gti.uvigo.es

ABSTRACT

We present our work to build the Spanish version of *SimpleNLG* by adapting it and creating new code to satisfy the Spanish linguistic requirements. Not only have we developed this version but also we have achieved a library that only needs the main words as input and it is able to conduct the generation process on its own. The adaptation of the library uses *aLexiS*, a complete and reliable lexicon with morphology that we created. On the other hand, our enhanced version uses *Elsa* created from the pictogram domain, which also contains syntactic and semantic information needed to conduct the generation process automatically. Both the adaptation and its enhanced version may be useful integrated in several applications as well as web applications, bringing them natural language generation functionalities. We provide a use case of the system focused on Augmentative and Alternative Communication and online video content services.

CCS Concepts

• **Human-centered computing** → User interface management systems.

Keywords

Natural Language Generation; *SimpleNLG*; Spanish; Alternative and Augmentative Communication; Online Video Content Services

1. INTRODUCTION

Recent studies indicate 91M of the world population [2] has been diagnosed with severe communication disorders. In children these

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

W4A '18, April 23–25, 2018, Lyon, France
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5651-0/18/04\$15.00
<https://doi.org/10.1145/3192714.3192837>

may be due to, e.g., Autism Spectrum Disorder (ASD) [17]. In Spain, according to a 2014 report¹, over 60,000 Spanish people had expression problems or communication disorders [7]. Despite of having language aids, these individuals are not enough motivated to communicate. These are the reasons why they need pictograms to interact. By minimizing attentional demands of people with ASD, requiring them to look only at a small spatial area (television or computer), they are more able to direct their focus to relevant stimuli [1] and stay motivated [11]. In most cases they are visual learners². On the other hand, multimedia content, as YouTube, is interesting because of the simultaneous use of various sensory channels, but also ubiquity for playful and educational purposes [13]. But, most of these videos are inaccessible to people with ASD, because their retrieval involves communication skills to make a query to these media that they do not always have. We plan to solve this problem using automatic natural language generation (NLG).

The paper is organized as follows: Section 2 reviews the related work; Section 3 explains the Spanish adaptation process and the *aLexiS* lexicon; Section 4 clarifies the Spanish enhanced version of *SimpleNLG*, including the *Elsa* lexicon; Section 5 presents a preliminary evaluation as well as a use case concerning online video content services; and Section 6 concludes the paper.

2. RELATED WORK

Some previous language aids on Augmentative and Alternative Communications (AAC) parse a telegraphic sequence and then re-generate a full sentence in natural language [4]. In it, the reconstruction of the meaning becomes difficult, because of the lack of morphological and syntactic structure, such as tools like *Talk Together* and *LetMe Talk*³. In [19], a Pictograph-to-text translation system for Dutch was presented. However, the Spanish version is still under development. More closely to our work, in [21] an NLG system based on highly domain dependent templates was described. The interest of NLG for ASD is illustrated by the several works from the state of the art [14, 15]. The interest of NLG applied to AAC using pictograms is also growing [6, 20].

¹ Press release available Oct. 2016 at http://www.dependencia.imserso.es/InterPresent2/groups/imserso/documents/binario/bde_pcd_2014.pdf

² Press releases available Oct. 2017 at <http://www.watchmelearn.com/video-modeling/video-modelingin-practice>.

Our challenge is to develop a system that generates sentences in natural language for Spanish providing a use case where the system assists people to formulate queries on an online video content service with the objective of retrieving a set of videos. To the best of our knowledge, there is no previous work focused on creating an AAC-specific NLG library for Spanish.

3. SPANISH SimpleNLG ADAPTATION

SimpleNLG library conducts the surface realization task with a knowledge-based approach. It is available for English, although there also exist versions for Brazilian Portuguese [5], French [23], German [3] and Italian [12]. A preliminary library for Spanish was presented [22] with no lexicon included and the NLG task was not performed automatically from the main words. A recently adaptation for Spanish [17] of this library was proposed but the differences are the automatic phase and the used lexicon. Besides, it has no connection with the AAC field.

The Spanish version of *SimpleNLG* was created starting from the French one, taking advantage of the linguistic proximity between both languages. The main adaptations are explained below, following the structure of separated modules containing rules for each processing level and language during the adaptation process. In this way, the adaptations were made for Spanish by adding syntactic and lexical features and re-encoding the corresponding rules in the French version within the grammar rules modules.

3.1 LEXICAL RESOURCE: *aLexiS*

Since the original English version of *SimpleNLG* and also all its adaptations to other languages use a small set of words (~6k words for the original English version and ~4k entries on the French lexicon) such as nouns, verbs and adjectives, among other lexical categories; we provide a complete and reliable lexicon, *aLexiS* [9], within the Spanish adapted version. This resource contains morphological information for nearly ninety thousand lemmas to increase the performance of the library in several contexts and applications.

4. ENHANCED SimpleNLG VERSION

In order to conduct Spanish NLG in a fully automatic way only requiring the users to provide the main word inputs, we applied a hybrid approach combining the linguistic knowledge given by *Elsa* with a language model, according to a statistical approach, with the aim of inferring prepositions. Our enhanced *SimpleNLG* version for Spanish makes use of this knowledge. Together with the lexical rules in the adapted library and those we implemented in the enhanced version, it can generate coherent and complete sentences from pictograms. More specifically it uses the textual representation of pictograms, keeping in mind that our target users are people diagnosed with communication disorders.

4.1 LEXICAL RESOURCE: *Elsa* LEXICON

Elsa was created from the pictogram domain and contains morphological, syntactic and semantic information needed to conduct the generation process automatically. We used *Arasaac*³, where each element and its meaning are linked to a single word or a sequence of words. Next, linguistic information was added to the *Arasaac* word entries. Keeping in mind that our target audience is AAC users, and in order to avoid pictograms related to prepositions, we developed a language model by means of a training process, considering bigrams and trigrams around verbs.

A more detailed explanation of the *Elsa* lexicon can be found in [8].

4.2 MORPHOLOGICAL RULES

Firstly, gender, number and person features for the whole clause derived from the words given by the user, have to be inferred. These characteristics are determined by the subject which is expected to be a nominal syntagm. We consider that it may be a pronoun, a proper noun or a noun. For example in the sentence *La niña es bonita* ‘The girl is beautiful’, the subject is a nominal syntagm composed of a feminine singular noun, *niña* ‘girl’ and a feminine singular determiner, *la* ‘the’. Consequently, the clause contains these features and remains in third person. This is why the verb *ser* ‘to be’ and the adjective *bonita* ‘beautiful’ are inflected in that way, this is in feminine, singular and third person. Once these features are inferred and considering that the lexicon contains the inflections of each lemma according to number and gender changes and, in the case of pronouns and verbs, also to person features, it is only necessary to apply these features to all word inputs. Nevertheless, sometimes there is no subject included and default features should be considered. In our case, first person, masculine gender, singular number and present tense.

If the user wants to add a complement with different features from the main clause or from the default ones (first person, masculine gender and singular number), it is necessary to introduce the inflected form, i.e., to introduce the pictogram with gender or/and number or/and person considerations. For example, if the user desires to generate the clause *La niña juega con las muñecas* ‘The girl plays with the dolls’; the word inputs should be *niña, jugar, muñecas* ‘girl, play, dolls’, where the word *muñecas* ‘dolls’ is in plural form, since there is a specific pictogram that represents this number consideration. Otherwise the resulting sentence would be *La niña juega con la muñeca* ‘The girl plays with the doll’.

4.3 SYNTACTIC RULES

The main challenge during the development of the Spanish *SimpleNLG* version was to infer the syntactic structure of the clause since we only had the word inputs in a sequential order. For this purpose, *Elsa* becomes crucial as it contains a wide range of linguistic data (morphological, syntactic and semantic information). We designed an algorithm able to infer the role of a word within the clause from its linguistic data and position within the clause. The syntactic rules of the algorithm are mostly based on the subcategorization frames of the Spanish verbs present in the lexicon. From this knowledge the algorithm is able to infer the syntactic structure of the final sentence. Our systems takes care of verbal constructions as well as compound subjects and Spanish double negation phenomenon.

4.4 SEMANTIC RULES

As mentioned before, our lexicon also contains semantic information for adjectives, adverbs, nouns and verbs. This semantic tagged data, together with the subcategorization frames for verbs, allows the system to generate coherent clauses. For example, from the words *yo, ir, parque, ayer* ‘I, go, park, yesterday’, the library is able to determine that *ir* ‘go’ is a verb followed by a noun which acts as an adverbial of place *parque* ‘park’. In addition, this verb needs the preposition a ‘to’ when the next given word is a place, semantically speaking, like *parque* ‘park’, according to the information in *Elsa*. Besides, the presence of the adverb *ayer* ‘yesterday’ indicates that the tense should be past (see Section 4.5). The resulting generated clause is *Yo fui al parque ayer* ‘I went to the park yesterday’. Without the information provided by *Elsa*, our system would not be able to

³ Available at <http://www.catedu.es/arasaac/>.

infer the additional elements needed, nor the correct morphological inflections related to the syntactic and semantic features. The resulting sentence would be *Yo voy el parque ayer* ‘I go the park yesterday’. The reason is that our NLG system is knowledge and statistically based and it requires linguistic data about the inputs to apply appropriate linguistic rules. We remark that when none of them can be applied, default rules are employed. As previously mentioned, our systems is able to create coordinate sentences.

4.5 DEFAULT RULES FOR AAC

A fully automatic Spanish NLG library requires default rules for atypical situations, such as using as system input words that are not included in the lexicon. In that case, the new given word is considered a proper noun. Let us recall again that our objective is to provide a fully automatic NLG system for people with communication disabilities, who cannot be trusted to introduce the linguistic information for new words. For example, when the user does not provide any subject, it is automatically considered to be *yo* ‘I’. The same occurs when no related features are given or the features cannot be inflected by the input words. Within this set of default rules are also included those which adjust the inflection of the verb according to the presence of a time-related adverb.

5. EXPERIMENTAL RESULTS

In order to provide a comprehensive evaluation of our adaptation of *SimpleNLG* to Spanish, we made the system generate sentences obtained from SimpleNLGENFr1.1. The tests covered various features of the Spanish grammar such as different types of sentences (affirmative, negative, interrogative, coordinate, passive, etc.), the whole verb conjugation for Spanish and constructions with different categories of words (adjectives, nouns, pronouns, etc.). Our adaptation passed the original tests successfully. On the other hand, given our interest in AAC and considering the fact that there are few sentences in both pictogram and Spanish natural language format to build a representative dataset, an external annotator⁴ was used to build a test set with over one hundred sentences⁵. After generating the sentences with the Spanish NLG library, we obtained a correctness precision (when the generated clause and the sentence from the test set are exactly the same) of approximately 77%.

In this regard, children and adult diagnosed with communication disorders will benefit from using the system by means of obtaining feedback related to what they generate. In this regard, this version was designed paying special attention to the storage and processing speed in order to make feasible its integration in a wide range of digital devices and indeed it was integrated within an Android application named *PictoDroid Lite* [10] which belongs to *Accegal*⁶. In this way, the users select a minimum set of pictograms through *PictoDroid Lite* and thanks to the embedded enhanced system, this sequence is transformed into a complete and coherent clause in natural language for Spanish. Then, a query

is created from this clause and finally the users get feedback in video support which is related to the topic they talk about. In Figures 1 and 2, we can observe the use case of the library within *PictoDroid Lite* related to the web. Firstly, in Figure 1, the interface of *PictoDroid Lite* can be observed. In this example, the selected pictograms are *lobo* ‘wolf’, *comer* ‘eat’ and *abuela* ‘grandmother’. The natural language sentence in Spanish associated with the selected pictograms is *El lobo come a la abuela* ‘The wolf eats the grandmother’ and this is what is searched in YouTube in Figure 2. In this regard, in Figure 2 is shown the video content related to the pictograms selected by the user, so he/she can consume online video content related to what he/she likes; taking into account how difficult is for people diagnosed with communication disorders to create sentences rather than using pictograms. Since the user only selects pictograms and then a complete and coherent sentence in natural language for Spanish is created, this may help him/her in the sense of learning how to verbalize thoughts and feelings.

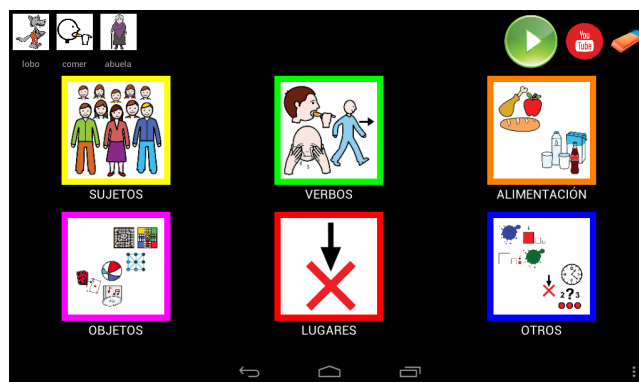


Figure 1 *PictoDroid Lite* Android application interface.

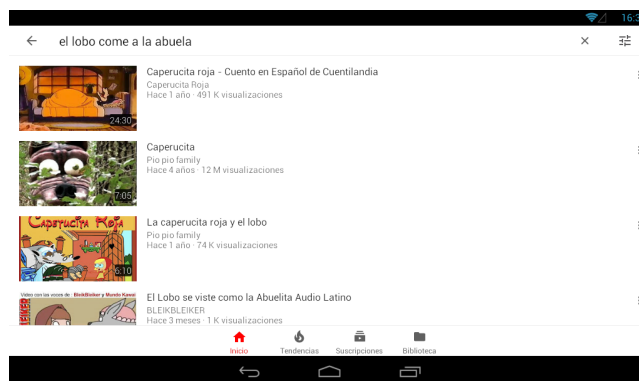


Figure 2 Results given by YouTube searching the natural language conversion of the pictograms selected by the user.

⁴ A researcher in the field of NLP and more concretely working on the NLG topic from the GTI Research Group in *atlantTic* - University of Vigo.

⁵ Available at <http://cuentosaulainfantil.blogspot.com.es/>, <http://es.scribd.com/doc/132908174/Oraciones-con-pictogramas-para-comprension-lectora#scribd>, <http://es.slideshare.net/nuriahs6/frases-en-pictogramas> and <http://www.slideshare.net/anabelcor/elena-juega-a-ser-sirena?player=js>.

⁶ <http://www.accegal.org/>. Available Oct. 2017

6. CONCLUSIONS

Both the adaptation of *SimpleNLG* to Spanish and the enhanced version represents a novel approach in Spanish NLG. The first can be used to give NLG functionalities to many applications. The second is especially useful for AAC and web applications. Spanish NLG and NLG applied to AAC are promising fields of study in which automation and flexibility deserve attention. Considering that the common practices in NLG are related to the use of templates in some steps of the generation process [16], our NLG

library may represent an alternative approach. We remark that our enhanced system, together with the Elsa lexicon, conducts the NLG task in a fully automatic way from a pictogram representation of the content (more specifically from a textual representation of the pictogram content, since we focus on text-to-text generation). To the best of our knowledge, there is not any similar tool for Spanish. Our NLG system has been especially designed paying attention to storage and processing speed to be compatible with a wide range of digital devices. Additional lexical resources can be easily integrated thanks to the automatic building process of both lexica.

7. ACKNOWLEDGMENTS

This work was partially supported by a grant from Ministerio de Economía, Industria y Competitividad, Spain (TEC2016-76465-C2-2-R); by a Xunta de Galicia grant (GRC2014/046); and by Red Temática en Tecnologías Emergentes de la Información y de las Comunicaciones (ED341D R2016/012).

8. REFERENCES

- [1] Albert Bandura. 1977. Social learning theory. Englewood Cliffs, N.J: Prentice Hall.
- [2] David Beukelman and Pat Mirenda. 2013. Augmentative and alternative communication: supporting children and adults with complex communication needs. Brookes Publishing.
- [3] Marcel Bollmann. 2011. Adapting *SimpleNLG* to German. In ENLG. ACL. Alessandro Mazzei, Cristina Battaglini, and Cristina Bosco. 2016. SimpleNLG-IT: adapting *SimpleNLG* to Italian. In INLG.
- [4] Ann Copestake. 1996. Applying Natural Language Processing Techniques to Speech Prostheses. Technical Report. Working Notes of the AAAI Fall Symposium on Developing Assistive Technology for People with Disabilities.
- [5] Rodrigo De Oliveira and Somayajulu Sripada. 2014. Adapting *SimpleNLG* for Brazilian Portuguese realisation. In INLG.
- [6] Chaohai Ding, Nawar Halabi, Lama Al-Zaben, Yunjia Li, EA Draffan, and Mike Wald. 2015. A web based multi-linguists symbol-to-text AAC application. In Proceedings of the 12th Web for All Conference. ACM, 24.
- [7] Fátima María García Doval. 2013. Aportaciones didácticas de un tablero digital para personas con dificultades de competencia comunicativa. Ph.D. Dissertation. University of Santiago de Compostela.
- [8] Silvia García-Méndez, Milagros Fernández-Gavilanes, Enrique Costa-Montenegro, Jonathan Juncal-Martínez, and Francisco J González-Castaño. 2017. Lexicon for Natural Language Generation in Spanish Adapted to Alternative and Augmentative Communication. LiRA-INLG (2017), 11–15.
- [9] Silvia García-Méndez, Milagros Fernández-Gavilanes, Enrique Costa-Montenegro, Jonathan Juncal-Martínez, and Francisco J González-Castaño. 2018. Natural Language Generation library for automatic Spanish text generation. Expert System With Applications (under review) (2018).
- [10] Jonathan Juncal-Martínez, Enrique Costa-Montenegro, Fátima Ma García-Doval, María Pousada-Carballo, and Ana Belén Barragáns-Martínez. 2015. ACCEGAL: Comunicación y estimulación cognitiva para dispositivos móviles Android. (2015).
- [11] Josefina Lozano Martínez, F Javier Ballesta Pagán, Salvador Alcaraz García, and M^a Carmen Cerezo Máiquez. 2013. Las tecnologías de la información y comunicación (TIC) en el proceso de enseñanza y aprendizaje del alumnado con trastorno del espectro autista (TEA). Revista Fuentes 14 (2013), 193–208.
- [12] Alessandro Mazzei, Cristina Battaglini, and Cristina Bosco. 2016. SimpleNLG-IT: adapting *SimpleNLG* to Italian. In INLG.
- [13] Luis Adiel Morales-Ramos and Teresa Guzmán-Flores. 2015. El vídeo como recurso didáctico para reforzar el conocimiento. (2015).
- [14] Paloma Moreda, Elena Lloret, et al. 2015. The use of Natural Language Processing techniques to support Health Literacy: an evidence-based review. (2015).
- [15] Christos Nikolopoulos, Deitra Kuester, Mark Sheehan, Shashwati Ramteke, Aniket Karmarkar, Supriya Thota, Joseph Kearney, Curtis Boirum, Sunnihith Bojedla, and Angela Lee. 2011. Robotic agents used to help teach social skills to children with autism: the third generation. In RO-MAN, 2011 IEEE. IEEE, 253–258.
- [16] A. Ramos-Soto, A. Bugarín, and S. Barro. 2016. On the Role of Linguistic Descriptions of Data in the Building of Natural Language Generation Systems. Fuzzy Sets Syst. 285, C (Feb. 2016), 31–51.
- [17] Alejandro Ramos-Soto, Julio Janeiro-Gallardo, and Alberto Bugarín. 2017. Adapting *SimpleNLG* to Spanish. In Proceedings of the 10th International Conference on Natural Language Generation (2017), 144–148.
- [18] Hospital Seattle-Children's. 2016. Augmentative and Alternative Communication. (2016). Available 20/10/2016 at www.seattlechildrens.org/pdf/PE1632S.pdf.
- [19] Leen Sevens, Vincent Vandeghinste, Ineke Schuurman, and Frank Van Eynde. 2015. Natural Language Generation from Pictographs. In ENLG. ACL, Brighton, UK.
- [20] Toshiyuki Takasaki and Yumiko Mori. 2007. Design and development of a pictogram communication system for children around the world. In Intercultural collaboration. Springer, 193–206.
- [21] Nava Tintarev, Ehud Reiter, Rolf Black, Annalu Waller, and Joe Reddington. 2016. Personal storytelling: Using Natural Language Generation for children with complex communication needs, in the wild. International Journal of Human-Computer Studies 92-93 (2016), 1–16.
- [22] Damian Jozef Trzpis. 2015. Adaptación de una herramienta de Generación de Lenguaje Natural al idioma Español. Ph.D. Dissertation. ETSI Informática. Universidad Politécnica de Madrid.
- [23] Pierre-Luc Vaudry and Guy Lapalme. 2013. Adapting *SimpleNLG* for Bilingual English-French Realisation. In ENLG. ACL, Sofia, Bulgaria.