Lexicon for Natural Language Generation in Spanish Adapted to Alternative and Augmentative Communication

Silvia García-Méndez, Milagros Fernández-Gavilanes, Enrique Costa-Montenegro, Jonathan Juncal-Martínez, Francisco J. González-Castaño
GTI Research Group, AtlantTIC
University of Vigo, 36310 Vigo, Spain
{sgarcia, mfgavilanes, kike, jonijm, javier}@gti.uvigo.es

Abstract

In this paper we present *Elsa*, the first lexicon for Spanish with morphological, syntactic and semantic information automatically generated from a well-known pictogram resource and especially tailored for Augmentative and Alternative Communication (AAC). This lexicon, focusing on that specific icon set widely used within AAC applications, is motivated by the need to improve Natural Language Generation (NLG) systems to aid people who have been diagnosed to suffer from communication disorders. In addition, we design an automatic lexicon extension procedure by means of a training process to complete the linguistic data. For this we used a dataset composed of novels and tales in Spanish, with pictogram representations, since the lexicon is meant for AAC applications for children with disabilities. Moreover, we provide the algorithms used to build our lexicon and a use case of *Elsa* within an NLG system to observe the usability of our proposal.

1 Introduction

According to the *State Database of Persons with Disabilities 2014* report[1], 14,456 Spanish people had expression problems, 72,088 had mixed disabilities and 45,818 had communication disorders (Doval, 2013). Relying on unofficial sources[2], in Spain and Mexico over 1% of children are autistic (over 800,000 people) requiring language aids. Many of these have evolved from graphical systems and rely on speech synthesis and speech recognition (Heimann Mühlenbock & Lundälv, 2011). They are known as Augmentative and Alternative Communication or AAC.

Our goal is to automatically create a Spanish vocabulary to be used in a Natural Language Generation (NLG) system applied to an AAC communicator, by merging different linguistic resources. Pictograms (used as input) act as a bridge between the lexicon and the NLG system, and help target users express themselves easily and quickly. Some previous AAC tools such as *Talk Together* or *LetMe Talk*[3] include small vocabulary packages with hand-coded knowledge, but none of them considers morphological, syntactic and semantic information when generating messages in Spanish. There is some work on language resource merging in the literature on manual and automatic management (Hughes, Souter, & Atwell, 1995; Crouch & King, 2005; Molinero, Sagot & Nicolas, 2009) but Spanish resources considering morphological, syntactic and semantic data has not been considered so far. Moreover, combining existing resources seemed a promising approach towards our goal, due to the grammatical difficulties of Spanish[4], as there are fewer resources than in English (Janssen, 2005).

The rest of this work is organised as follows. In Section 2 we review the existing linguistic resources for Spanish and the process conducted to build *Elsa*. In Section 3 we present an automatic lexicon extension procedure. Then, in Section 4 we conduct an

[1] Press release available Oct. 2016 at http://www.dependencia.imserso.es/InterPresent2/groups/imserso/documents/binario/bdepcd_2014.pdf.

[2] Press releases available Oct. 2016 at http://www.antena3.com/noticias/salud/espana-350000-personas-estan-diagnosticadas-autismo_20150402571edda86584a8abb583c1ee.html and http://projectspectrum.net/2016/02/05/primer-estimado-de-prevalencia-de-autismo-en-mexico-es-de-1-en-115-individuos.

[3] Available at http://acecentre.org.uk/talk-together and http://www.utac.cat/descarregues/cace-utac.

[4] Some difficulties are those related to the inflection of verbs.

Proceedings of the Linguistic Resources for Automatic Natural Language Generation Workshop, pages 11–15, Santiago de Compostela, Spain, September 4, 2017.

evaluation of the created lexicon. In Section 5 we provide a use case of *Elsa* within an NLG system. Section 6 concludes the paper.

2 Reusing existing resources to build *Elsa*

The construction of *Elsa* begins with the selection of an available pictographic set for AAC users. After evaluating some possibilities[5] we choose the free and highly comprehensive *Arasaac*[6] set. Nonetheless, this dataset had to be preprocessed by removing the pictograms with the same meaning and word descriptions, whose redundancy is due to their different representation according to their colour. By doing so, we obtained our icon set with 9,411 pictograms, of which 6,970 have a single associated word (including proper names) and the rest are verbal phrases or compound proper names. Once this step is finished, it is necessary to add POS tags, syntactic and semantic information to each *Arasaac* word entry. For this purpose, we looked for available Spanish linguistic resources. We choose:

- *Adesse*[7] (García-Miguel, Vaamonde, & González Domínguez, 2010).
- Multilingual Central Repository[8] (MCR) (González-Agirre, Laparra, & Rigau, 2012).
- Lexicon of Spanish inflected forms[9] (LEFFE) (Molinero, Sagot, & Nicolas, 2009).

We start the process by extracting from each resource some information on the forms of the preprocessed icon set. Next, our approach follows the two well-defined steps (Crouch & King, 2005) between which we include a verification step: (1) we extract and map the form entries to a common format, adapted from Lexical Markup Framework (LMF) format (Francopoulo, et al., 2006); (2) we verify them at a lexical level in the DRAE[10]; and finally (3) we combine the entries once they have been found to be equivalent using the graph unification model (Necsulescu, Bel, Padró, Marimon & Revilla, 2011; Bel, Padró & Necsulescu, 2011). This operation is based on set unions of compatible feature values, allowing the validation of common information, the addition of differential information and the exclusion of inconsistencies.

The steps of extraction and mapping, verification and merging are explained in Algorithms 1, 2 and 3, respectively. In algorithm 4, we can observe the composition of the steps as explained in this Section.

Algorithm 1 Extraction and mapping

```
function EXTRACTION_MAPPING({LEFFE})
    for e_LEFFE ∈ {LEFFE} do
        lem_eLEFFE = e_LEFFE.getLemma()
        cat_eLEFFE = e_LEFFE.getCat()
        if cat_eLEFFE=verb AND lem_eLEFFE.isInAdesse()
        then
            e_ADESSE = searchInAdesse(lem_eLEFFE)
            {ADESSE}.add(e_ADESSE)
        end if
        if cat.isAdj() OR cat.isAdv() OR cat.isN() OR
         cat.isV() AND lem_eLEFFE.isInMCR() then
            e_MCR = searchInMcr(lem_eLEFFE)
            {MCR}.add(e_MCR)
        end if
    end for
end function
```

Algorithm 2 Verification

```
function VERIFICATION({SET})
    for e_SET ∈ {SET} do
        lem_eSET = e_SET.getLemma()
        cat_eSET = e_SET.getCat()
        if !lem_eSET.isInDRAE() OR !lem_eSET.catInDRAE
(cat_eSET)
        then
            {SET}.delete(e_SET)
        end if
    end for
end function
```

[5] Some of them were: *Pictographic Communication System* (http://www.mayer-johnson.com/category/symbols-and-photos) and *Pictogram* (http://www.pictogram.se), which are not free; or *Widgit* (https://widgit.com) with no support for Spanish.

[6] Created by the CATEDU, the *Alborada Special Education Public School* and Sergio Palao in 2008 under Creative Commons license. It contains over 16,000 pictograms with their associated words or sequence of words for Spanish, as well as multiple other languages. Available at http://www.catedu.es/arasaac.

[7] Database of over 3,400 verbs, diathesis alternations and syntactic semantic schemes in Spanish. Accessible at http://adesse.uvigo.es, May 2017.

[8] Lexical database integrating the Spanish WordNet into the EuroWordNet framework. Available at http://adimen.si.ehu.es/web/MCR, May 2017.

[9] A wide-coverage morphological and syntactic lexicon. Available at https://gforge.inria.fr/frs/?group%20id=482&release%20id=4290, May 2017.

[10] *Diccionario de la Real Academia de la Lengua Española* available at http://www.rae.es.

Algorithm 3 Merging

function MERGING
{ELSA} = {LEFFE} ∪ ∪ {ADESSE} ∪ {MCR}
end function

Algorithm 4 Building procedure

{ADESSE} = {∅}, {MCR} = {∅}
{LEFFE}= LoadLeffe()
EXTRACTION_MAPPING({LEFFE})
{SETS} = {{LEFFE},{ADESSE},{MCR}}
for {set} ∈ {SETS} **do**
VERIFICATION({set})
end for
MERGING

3 Automatic lexicon extension

Keeping in mind that we intend to use this lexicon within an NLG system adapted to AAC users, and in order to facilitate the task of avoiding pictograms related to prepositions, we need to infer *a priori* which specific preposition follows a verb. The training process was performed using a dataset composed of novels and nearly five hundred tales in Spanish (Andersen, 2016; Anonymous, 2016; Grimm, 2016), previously POS-tagged applied with Freeling Tagger[11], since we plan to use the lexicon in AAC applications for children with disabilities and these are the only contents with pictogram representation. In this regard, we are able to include more options beyond those present in the subcategorization frames for verbs taken from LEFFE and *Adesse*, such as those related to figurative language approaches[12]. Since this grammar realization is not present in the selected lexica, we developed a language model from a training process, considering bigrams and trigrams around verbs and using syntactic and semantic knowledge.

4 Experimental results

In order to evaluate the quality of *Elsa*, we first measured the coverage achieved after adding the information extracted from all resources. Table 1 shows the number of lemmas that were included in each resource. Table 2 shows the coverage of *Elsa* over the icon set. Our lexicon covered almost the entire icon set and most word entries include syntactic and semantic data essential to conduct the NLG process correctly. Moreover, Table 3 shows the number of lemmas and forms classified by categories. Most lemmas (3,165) are tagged as nouns, representing 7,035 inflected forms added to *Elsa*, whereas most forms (45,341) are tagged as verbs, representing 811 lemmas.

```
<Entry lemma="desagradar">
 <fileName>desagradar.png</fileName>
 <feat att="POS" val="v"/>
 <SemSynset>
  <feat att="id" val="ili3001776727"/>
  <feat att="genSem" val="emotion"/>
  <feat att="concretSem"
   val="IntentionalPsychologicalProcess"/>
 </SemSynset>
 <WordForm>
  <feat att="writtenForm" val="desagradar"/>
  <feat att="pronominal" val="true"/>
  <feat att="transitive" val="false"/>
  <feat att="gerund" val="desagradando"/>
  <feat att="p_ms" val="desagradado"/>
  <feat att="p_mp" val="desagradados"/>
  <feat att="p_fs" val="desagradada"/>
  <feat att="p_fp" val="desagradadas"/>
  <feat att="IP1s" val="desagrado"/>
  ...
 </WordForm>
 <SCF>
  <feat att="type" val="active"/>
  <feat att="subj" val="est"/>
  <feat att="oind" val="exp"/>
 </SCF>
 <SCF_training>
  <feat att="type" val="active"/>
  <feat att="subj" val="est"/>
  <feat att="a_oind" val="exp"/>
 </SCF_training>
</Entry>
```

Figure 1: Fragment of the entry for the Spanish lemma *desagradar* 'displease' in *Elsa*

Figure 1 shows an example in adapted LMF format of the word entry *desagradar* 'displease' with morphological, syntactic and semantic data. *SemSynset* contains the semantic information from MCR. In addition to the conjugation in *WordForm*, *SCF* contains the syntactic information on the verb after combining the data extracted from LEFFE and *Adesse*. There is only one possible realization in the active form, where the subject is an *estímulo* 'stim-

[11] A library that provides multiple language analysis services, including probabilistic prediction of categories in unknown words (Atserias et al., 2006; Padró & Stanilovsky, 2012).

[12] For example the verb *comer* 'eat' whose subcategorization frames did not include the possibility of using the preposition *a* with people, even though it is widely employed in tales, as in the clause *El lobo se comío a la abuelita* 'The wolf ate the granny'.

Category	LEFFE		*Adesse*		MCR	
	Number	%	Number	%	Number	%
Adjective	824	98.92%	-	-	503	60.38%
Adverb	59	100.00%	-	-	46	77.97%
Noun	3,129	98.86%	-	-	2,957	93.43%
Verb	809	99.75%	731	90.14%	791	97.53%

Table 1: Lemmas of *Elsa* included in the different resources by lexical category

ulus' and the verb is followed by an indirect object[13] that is an *experimentador* 'experimenter'.

In *SCF_training*, a new preposition, *a* 'to' (not present in any of the selected resources), was inferred in order to use it within the subcategorization frame of the verb (before the indirect object `oind`). In addition, the *Elsa* entry *desagradar* 'displease' is linked to a pictogram image file from the icon set.

Included	Single word	Compound
in *Elsa*	4,434	684
not in *Elsa*	1,716	1,167
TOTAL	**6,150**	**1,851**

Table 2: *Elsa* coverage of the AAC icon set

	Category	Lemmas	Forms
Elsa	Adjective	833	2,583
	Adverb	59	59
	Conjunction	15	15
	Determiner	23	71
	Noun	3,165	7,035
	Preposition	20	20
	Pronoun	16	47
	Proper name	171	171
	Verb	811	45,341
	TOTAL	**5,298**	**55,342**

Table 3: *Elsa* size by category

5 *Elsa* use case: NLG system

Assuming that our system input is: *tiempo, desagradar, profesor, ayer* 'weather, displease, teacher, yesterday'.

- System output **using *Elsa***: *El tiempo desagradó al profesor ayer* 'The weather displeased the teacher yesterday'.
- System output without **using *Elsa***: *El tiempo desagradar el profesor ayer* 'The weather displease the teacher yesterday' (where displease is the infinitive of the verb).

In this example, the system can determine that *desagradar* 'displease' is a verb, whose subject is *tiempo* 'weather', followed by an indirect object *profesor* 'teacher'. In addition, this verb needs the preposition *a* 'to' because *profesor* 'teacher' is a person. Besides, the presence of the adverb *ayer* 'yesterday' indicates that the tense is past. Our system would neither infer the additional elements needed nor the correct morphological inflections related to the syntactic and semantic features without the linguistic information provided by *Elsa*.

6 Conclusions

Elsa is an approach for lexica generation specially tailored for the needs of AAC applications. Besides including several types of linguistic information (morphology, syntax and semantics), a training process was executed to complete the subcategorization frames for verbs, like those present in figurative language. The resulting lexicon may be useful for assisting people with communication disabilities through NLG systems. In order to increase efficiency and precision, additional linguistic resources can be easily integrated due to the fact that the building process is automatic. To complete the semantic information, we propose to establish synonymy relations between the word entries to reuse their semantic classification and fill in the missing information.

[13] In Spanish an intransitive verb has not direct object but it may be followed by an indirect object or other complements.

Acknowledgments

This work was partially supported by a grant from Ministerio de Economía, Industria y Competitividad, Spain (TEC2016-76465-C2-2-R); and by Xunta de Galicia grant (GRC2014/046).

References

Andersen. 2016. *Fairy tales of Hans Christian Andersen* (Spanish). Available 28/11/2016 at https://es.wikisource.org/wiki/Cuentos_clásicos_para_niños.

Anonymous. 2016. *Traditional Spanish tales*. Available 28/11/2016 at http://loscuentostradicionales.blogspot.com.es.

Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source NLP library. In *Proceedings of the fifth international conference on Language Resources and Evaluation* (LREC 2006), Genoa, Italy, ELRA, pp 48-55.

Nuria Bel, Muntsa Padró, and Silvia Necsulescu. 2011. A method towards the fully automatic merging of lexical resources. In *Proceedings of Workshop on Language Resources, Technology and Services in the Sharing Paradigm*. ACL, Chiang Mai, Thailand, pp 8-15.

Dick Crouch and Tracy Holloway King. 2005. Unifying lexical resources. In *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*. Saarbrücken, Germany, pp 32-37.

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2006. Lexical markup framework (LMF) for NLP multilingual resources. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, MLRI '06, Stroudsburg, PA, USA. Association for Computational Linguistics, pp 1-8.

Fátima María García Doval. 2013. *Aportaciones didácticas de un tablero digital para personas con dificultades de competencia comunicativa.* Ph.D. thesis, University of Santiago de Compostela.

José M. García-Miguel, Gael Vaamonde, and Fita González Domínguez. 2010. Adesse, a database with syntactic and semantic annotation of a corpus of Spanish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation* (LREC'10) (eds. N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias), Valletta, Malta. European Language Resources Association (ELRA), pp 1903-1910.

Aitor González-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *6th Global WordNet Conference*, Matsue, Japan, pp 118-125.

Grimm. 2016. *Grimm's fairy tales* (Spanish). Available 28/11/2016 at http://www.grimmstories.com/es.

Katarina Heimann Mühlenbock and Mats Lundälv. 2011. *Using lexical and corpus resources for augmenting the AAC lexicon*. In Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, SLPAT '11, Edinburgh, Scotland, UK. Association for Computational Linguistics, pp 120-127.

John Hughes, Clive Souter, and Eric Atwell. 1995. Automatic extraction of tagset mappings from parallel annotated corpora. In *From Texts to Tags: Issues in Multilingual Language Analysis. Proceedings of SIGDAT Workshop in Conjunction with the 7h Conference of the European Chapter of the Association for Computational Linguistics*. University College Dublin, Ireland, pp 10-17.

Maarten Janssen. 2005. Open source lexical information network. In *Proceedings of the 3rd International Workshop on Generative Approaches to the Lexicon*, Geneva, Switzerland, pp 400-410. http://ww.cibercursoslp.com/Papers/GL2005-mjanssen.pdf. Date Accessed: April 18, 2017.

Miguel A. Molinero, Benoît Sagot, and Lionel Nicolas. 2009. A morphological and syntactic wide-coverage lexicon for Spanish: The Leffe. In *RANLP 2009 - Recent Advances in Natural Language Processing*, Borovets, Bulgaria, September, pp 264-269.

Silvia Necsulescu, Nuria Bel, Muntsa Padró, Montserra Marimon, and Eva Revilla. 2011. Towards the automatic merging of language resources. In *Proceedings of 1st International Workshop on Lexical Resources: an ESSLLI 2011 Workshop*, Ljubljana, Slovenia, pp 7-77.

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. ELRA, pp 2473-2479.