Teaching FORGe to Verbalize DBpedia Properties in Spanish

Simon Mille
Universitat Pompeu Fabra
Barcelona, Spain
simon.mille@upf.edu

Stamatia Dasiopoulou
Independent Researcher
Barcelona, Spain
stamatia.dasiopoulou@gmail.com

Beatriz Fisas
Universitat Pompeu Fabra
Barcelona, Spain
beatriz.fisas@upf.edu

Leo Wanner
ICREA and Universitat Pompeu Fabra
Barcelona, Spain
leo.wanner@upf.edu

Abstract

Statistical generators increasingly dominate the research in NLG. However, grammar-based generators that are grounded in a solid linguistic framework remain very competitive, especially for generation from deep knowledge structures. Furthermore, if built modularly, they can be ported to other genres and languages with a limited amount of work, without the need of the annotation of a considerable amount of training data. One of these generators is FORGe, which is based on the Meaning-Text Model. In the recent WebNLG challenge (the first comprehensive task addressing the mapping of RDF triples to text) FORGe ranked first with respect to the overall quality in human evaluation. We extend the coverage of FORGE's open source grammatical and lexical resources for English, so as to further improve the English outcome, and port them to Spanish, to achieve a comparable quality. This confirms that, as already observed in the case of SimpleNLG, a robust universal grammar-driven framework and a systematic organization of the linguistic resources can be an adequate choice for NLG applications.

1 Introduction

The origins of Natural Language Generation (NLG) are in rule-based sentence/text generation from numerical data or deep semantic structures. With the availability of large scale syntactically annotated corpora and the lack of publicly available knowledge repositories, the focus had shifted to statistical surface generation. However, thanks to Semantic Web (SW) initiatives such as the *W3C Linking Open Data Project*,[1] a tremendous amount of structured knowledge has been made publicly available as language-independent triples; the Linked Open Data (LOD) cloud currently contains over one thousand interlinked datasets (e.g., DBpedia, Wikidata), which cover a large range of domains and amount to billions of different triples. The verbalization of LOD triples, i.e., their mapping onto sentences in natural languages, has been attracting a growing interest in the past years, as shown by the organization of dedicated events such as the WebNLG 2016 workshop (Gardent and Gangemi, 2016) and the 2017 WebNLG challenge (Gardent et al., 2017b). As a result, a variety of new NLG systems designed specifically for handling structured data have emerged, most of them statistical, as seen in the 2017 WebNLG challenge, although a number of rule-based generators have also been presented. All systems focus on English, mainly because no training data other than for English are available as yet. Given the high cost for the creation of training data, this state of affairs is likely to persist for some time. Therefore, the question on the competitiveness of rule-based generators arises.

One of the rule-based generators presented at WebNLG was FORGe (Mille and Dasiopoulou, 2017), which ranked first with respect to overall quality in the human evaluation. FORGe is grounded in the linguistic model of the Meaning-Text Theory (Mel'čuk, 1988). The multistratal nature of this model allows for a modular organization of blocks of graph-transduction rules, from blocks that are universal, i.e., multilingual, to blocks that are language-specific. The graph-transduction framework MATE (Bohnet and Wanner, 2010) furthermore facilitates a systematic hierarchical rule writing and testing. SimpleNLG

[1] https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

Proceedings of The 12th International Conference on Natural Language Generation, pages 473–483,
Tokyo, Japan, 28 Oct - 1 Nov, 2019.

(Gatt and Reiter, 2009) demonstrated that a well-defined generation infrastructure, along with a transparent, easy to handle rule and structure format, is a key for its take up and use for creation of generation modules for multiple languages. In what follows, we aim to demonstrate that the FORGe generator can also well serve as a multilingual portable text generator for verbalization of structured data and that its lexical and grammatical resources can be easily extended to reach a higher coverage of linguistic constructions. For this, we extend its publicly available resources for English, so as to improve the quality of the English texts and port the resources to Spanish with a comparable output quality.

In the next section, we summarize the related work. Section 3 introduces FORGe. In Section 4, we outline our work on the extension of the available English resources and on the adaptation of FORGe to Spanish. Section 5 presents the results of the automatic evaluation of the extended system, and Section 6 a qualitative evaluation of the outputs in both languages. Section 7, finally, draws some conclusions and presents the future work.

2 Related work

The most prominent recent illustration of the portability of a generation framework is SimpleNLG. Originally developed for generation of English in practical applications (Gatt and Reiter, 2009), in the meantime it has been ported to generate, among others, in Brasilian Portuguese (De Oliveira and Sripada, 2014), Dutch (de Jong and Theune, 2018), German (Bollmann, 2011), Italian (Mazzei et al., 2016), and Spanish (Soto et al., 2017). However, while SimpleNLG is a framework for surface generation, usually with a limited coverage, we are interested in a portable multilingual framework for large scale text generation from structured data, more precisely, from DBpedia properties (Lehmann et al., 2015).

Although most existing NLG generators combine different techniques, there are three main approaches to generating texts from an input sequence of structured data (Bouayad-Agha et al., 2014; Gatt and Krahmer, 2018): (i) filling slot values in predefined sentence templates (Androutsopoulos et al., 2013), (ii) applying grammars (rules) that encode different types of linguistic knowledge (Wanner et al., 2010), and (iii) predicting statistically the most appropriate output (Gardent et al., 2017b; Belz et al., 2011). Template-based systems are very robust, but also limited in terms of portability since new templates need to be defined for every new domain, style, language, etc. Statistical systems have the best coverage, but the relevance and the quality of the produced texts cannot be ensured. Furthermore, they are fully dependent on the available (still scarce and mostly monolingual) training data. The development of grammar-based systems is time-consuming and they usually have coverage issues. However, they do not require training material, allow for a greater control over the outputs (e.g. for mitigating errors or tuning the output to a desired style), and the linguistic knowledge used for one domain or language can be reused for other domains and languages. In addition to these, a number of systems actually address the whole sequence as one step, by combining approaches (i) and (iii) and filling the slot values of pre-existing templates using neural network techniques (Nayak et al., 2017).

In the WebNLG challenge (Gardent et al., 2017a), systems of types (ii) and (iii) have been presented. The task consisted in generating texts from up to 7 DBpedia triples from 15 categories, covering in total 373 distinct DBpedia properties. Nine categories appeared in the training data ('Astronaut', 'Building', 'University', etc.), i.e., were "seen", and five categories were "unseen", i.e., they did not appear in the training data ('Athlete', 'Artist', etc.). At the time of the challenge, the WebNLG dataset contained about 10K distinct inputs and 25K data-text pairs; a sample data-text pair is shown in Figure 1. The neural generator ADAPT (Elder et al., 2018) performed best on seen data, and FORGe on unseen data and overall. In what follows, we aim to improve the performance of FORGe on seen data for English and furthermore port it to Spanish.

3 Overview of FORGe

FORGe is an open-source generator implemented in terms of graph transducers; it covers the last two typical NLG tasks (text planning and linguistic generation). Following the Meaning-Text Theory (Mel'čuk, 1988), FORGe is based on the notion of linguistic dependencies, that is, the semantic, syntactic and morphological relations between the components of the sentence. Input predicate-argument structures are mapped onto sentences by

<originaltripleset>
<otriple> Antwerp_International_Airport | city | Antwerp </otriple>
<otriple> Belgium | leader | Charles_Michel </otriple>
<otriple> Antwerp| country | Belgium </otriple>
<otriple> Belgium| language| German </otriple>
</originaltripleset>

Reference 1: Charles Michel is the leader of Belgium where the German language is spoken. Antwerp is located in the country and served by Antwerp International airport.
Reference 2: Antwerp International Airport serves the city of Antwerp which is a popular tourist destination in Belgium. One of the languages spoken in Belgium is German, and the leader is Charles Michel.

Figure 1: Sample pair of data (subject-property-object) and human-produced texts (*references*).

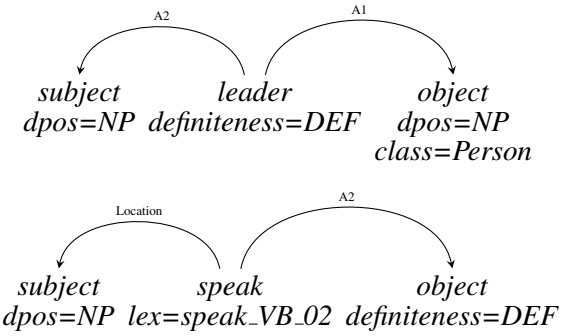

Figure 2: Sample PredArg templates corresponding to the *leader* (top) and *language* (bottom) properties.

applying a series of rule-based graph transducers. The generator handles Semantic Web inputs by means of introducing abstract predicate-argument (PredArg) templates and micro-planning grammars before the core linguistic generation module (Mille and Dasiopoulou, 2017).

3.1 Mapping properties to PredArg templates

Predicate-argument templates in a PropBank (Kingsbury and Palmer, 2002; Babko-Malaya, 2005) fashion were defined taking into account the property as well as the type of the subject and object values.[2] Thus, each of the properties found in the evaluation triples was associated to one of these templates. Parts of speech (e.g., NP –proper noun), grammatical features (e.g., verbal tense or nominal definiteness), or information from DBpedia (e.g., classes), for instance, can be specified in the template.[3] Figure 2 shows sample PredArg templates for the DBpedia properties *leader* and *language* respectively;[4] 318 templates were used for the 373 properties of WebNLG.

3.2 Population of the templates

Using the aforementioned mappings, each input triple is transformed into a respective PredArg structure. This involves two main steps. First, the cleaning of the object, including the extraction of value/unit information from datatype fillers and distinct values from list-like fillers. Second, if different from the template, the assignment of pertinent subject/object class labels, which are geared to the subsequent linguistic generation steps and currently include 'Person', 'Location', 'Time' (further distinguishing between date, year, month), and 'Literal' (i.e. datatype values). During this step, cardinality and number information labels are also assigned. Last, in the case of multiple triple inputs, the triples are ordered (as a preliminary step for the subsequent aggregation) based on the number of appearances of their subjects and on whether a subject of a triple serves also as an object in another triple. For the population of the templates of Figure 2, the subject and object placeholders are simply replaced by the corresponding subjects and objects of Figure 1, without cleaning or further modification.

3.3 Aggregation of PredArg structures

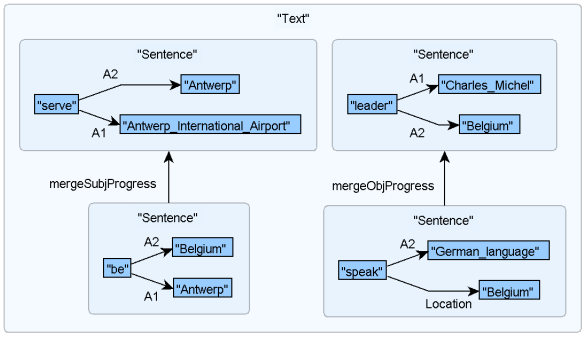

Figure 3: Aggregation of populated templates (step 2)

In order to group triples into complex sentences, a graph-transduction module that performs aggregation in two steps was developed. First, shared predicate–object argument pairs in the populated templates are targeted: if the object arguments have the same relation with their respective predicates, they will be coordinated (e.g., *Jazz*$_S$

[2]Inspection of subject and object types was needed as some properties denoted more than one meaning and corresponded to different templates.

[3]Unspecified values are assigned later in the process.

[4]Note that is possible to refer to a particular PropBank class in the PredArg graphs as, e.g., *speak_VB_02*, which corresponds to the second meaning of *speak* in ProbBank; if no class is indicated (e.g. *leader*), the first PropBank sense is assigned by default.

influenced$_P$ *funk*$_{O1}$ *and afrobeat*$_{O2}$.); if the relations are different, the objects become siblings under the first occurrence of the predicate (e.g. *[Alan Bean]*$_S$ *[was born]*$_P$ *[in Wheeler (Texas)]*$_{O1}$ *[on March 15]*$_{O2}$.); the duplicated nodes are removed. What is targeted in the second place is an argument of a predicate that appears further down in the ordered list of PredArg structures. If identified, the PredArg structures are merged by fusing the common argument; see e.g. Antwerp and Belgium in Figure 3, which are merged at the end of the process, c.f. Figure 4. During linguistic generation, this results in the introduction of post-nominal modifiers such as relative and participial clauses or appositions (see next section). In order to avoid the formation of heavy nominal groups, at most one aggregation is allowed per argument.

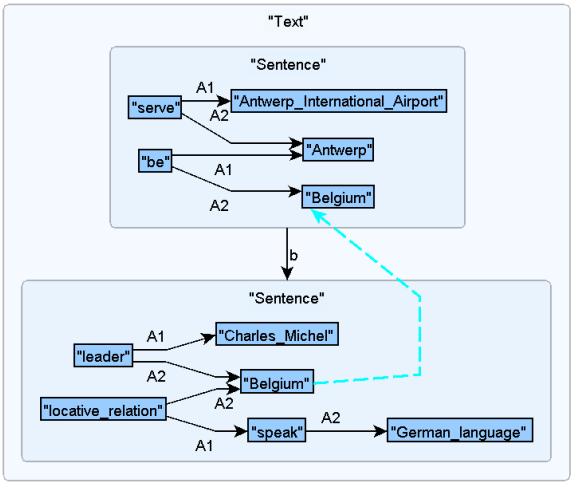

Figure 4: Aggregated PredArg structures

3.4 Linguistic generation

The next and last step is the rendering of the aggregated PredArg structures into sentences. This part of the system performs the following actions: (i) syntacticization of predicate-argument graphs; (ii) introduction of function words; (iii) linearization and retrieval of surface forms. First, a deep-syntactic (DSynt) structure is generated: missing parts of speech are assigned, the syntactic root of the sentence is chosen, and from there a syntactic tree over content words is built node by node; see Figure 5.[5] Then, as shown in Figure 6, functional words (prepositions, auxiliaries, determiners, etc.) are introduced and fine-grained surface-syntactic (SSynt) labels are established, using a subcategorisation lexicon. For this purpose, lexical resources derived from PropBank (Kingsbury and Palmer, 2002), NomBank (Meyers et al., 2004) or VerbNet (Schuler, 2005) are used; see (Mille and Wanner, 2015; Lareau et al., 2018). Personal and relative pronouns are introduced using the coreference relations (dotted arrows) and the *class* feature, which allows for distinguishing between human and non-human antecedents. Finally, morpho-syntactic agreements are resolved, the syntactic tree is linearized through the ordering of (i) governor/dependent and (ii) dependents with each other, and the surface forms are retrieved. Post-processing rules are then applied: upper casing, replacement of underscores by spaces, etc.

[5] Note that the node brought together during the previous step are not necessarily split up at this level.

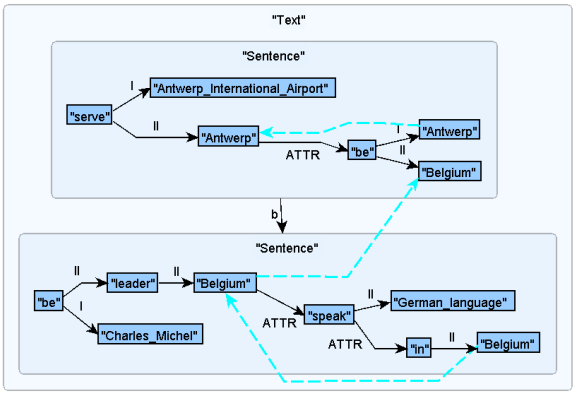

Figure 5: Deep-syntactic structures

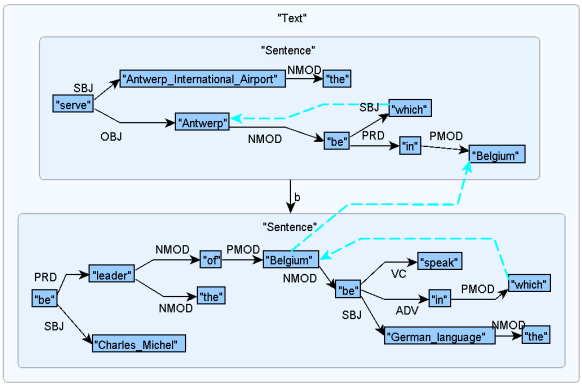

Figure 6: Surface-syntactic structures

Consider for instance the *leader* property of Figure 2 and selected phenomena: (i) the support verb *be* is established as the root (Fig. 5), (ii) the preposition *of* is introduced below *leader* (Fig. 6), and the *SBJ* relation is introduced between *be* and *Charles_Michel*, which (iii) causes the verb to be placed after the noun and get morphological agreement features from it (third person singular), while *NMOD* towards a preposi-

tion causes the opposite order and no agreement, etc.: *Charles_Michel*$_{3sg}$ > *is*$_{3sg}$ > *the* > *leader* >*of* >*Belgium*. The final sentence generated for the four triples is *The Antwerp International Airport serves Antwerp, which is in Belgium. Charles Michel is the leader of Belgium, in which the German language is spoken*.

4 Multilingual extension of FORGe

FORGe was developed primarily for English generation, and in order to port it to Spanish, four main aspects had to be addressed: (i) the PredArg templates, (ii) the grammatical resources, (iii) the lexical resources, and (iv) the translation of the Subject and Object values from the original DBpedia triples, in which English is used.

4.1 Adaptation of the PredArg templates

130 out of the 217 templates that cover all the WebNLG seen dataset were left unchanged, and 87 of them had to be adapted to Spanish (all templates stay in English). The adaptation of the templates consisted of two major modifications: on the one hand, some predicates were changed, such as the English predicate *parent_company*, which was modified to *daughter_company*, more idiomatic in Spanish (28 cases); and on the other hand, some predicate-argument relations have been updated in order to match the entries in the Spanish lexicon (50 cases).[6] The other modifications include: a change in the definiteness of a noun (7 cases, e.g. *chickens*$_{INDEF}$ in *Chickens belong to the class "bird"* would be rendered as *el*$_{DEF}$ *pollo* in Spanish) or the tense or aspect of a verb (6 cases), or the addition of a new predicate (1 case)[7].

4.2 Adapting the rules to Spanish

The most important rules added for Spanish are (i) rules introducing the surface-syntactic relations, based on which linear order and morphological agreements are resolved, (ii) rules for gender and number agreements in noun groups and auxiliary constructions, and (iii) word ordering rules. Note that the rules for Spanish also apply to other Romance languages with similar features (e.g. French, Italian, etc.).

For designing the rules, we followed the approach of AnCora-UPF (Mille et al., 2013), a Spanish dataset in which each dependency relation is associated with a set of syntactic properties. For instance, a *subject* is characterized by being linearized to the left of its governing verb (by default), by being removable, by triggering the number and person agreements on the verb, etc. During the linguistic generation stage, 27 out of the 47 relations proposed in AnCora-UPF [8] are currently supported.

In order to generalize the ordering rules across languages, the dependencies were introduced in the lexicon with details about how they are linearized with respect to their governor (*vertical* ordering). Generic linearization rules also apply. For instance, for the *copul* dependency (such as between *be* and *retired*), pronominal dependents are linearized BEFORE the finite verb, and the other dependents AFTER it. If several dependents end up at the same height with respect to their governor, they need to be ordered with each other. 21 rules were added to manage these *horizontal* orderings. They facilitate the ordering of, for instance, determiners before the adjectives, or small adverbial groups before the objects. Finally, 18 rules for resolving the agreements between verb and subject, adjective/determiner and noun, copulatives and subjects, etc. were implemented. For instance, in the structure *Joana*$_{3-FEM-SING}$ ←subj *estar* copul → *jubilado*, will be linearized and inflected as follows: *Juana está jubilada* (lit. 'Joana is$_{3-SING}$ retired$_{FEM-SING}$').

4.3 Crafting the Spanish dictionaries

Several types of dictionaries are needed for generation: (i) a dictionary that maps the input meanings/concepts onto lexical units of a particular language (called *concepticon*), (ii) a dictionary that contains the combinatorial properties of each lexical unit (*lexicon*), (iii) a dictionary with the full forms of the words (called *morphologicon*). Some other information, such as linearization properties of dependencies (see Section 4.2) are also better stored in the lexicon in order to allow for more generic (hence less numerous) rules.

As explained in Section 3.1, the DBpedia properties are mapped to PredArg structures. For the

[6] The mismatch between the relations in English and Spanish is simply practical; it originates from a detail in the implementation of the mapping of the arguments in both languages, so a unification of the rule design would solve it.

[7] More than one change can happen in one template.

[8] adjunct, adv, agent, analyt_fut, analyt_pass, analyt_perf, analyt_progr, aux_phras, appos, attr, compar, coord, coord_conj, copul, det, dobj, iobj, modal, modif, obl_compl, obl_obj, prepos, punc, quant, relat, sub_conj, and subj.

WebNLG challenge, English was the only language to generate, so the labels of the nodes in the PredArg templates were in English. In order to take advantage of the templates developed for FORGe in 2017, we also use these structures with English vocabulary as input to the generator. Thus, we manually crafted the concepticon (255 entries), in which the keys are the predicates from the templates, and the values are lexical units in Spanish; for instance, the predicate *locate* is mapped to the Spanish verb *estar_VB_04* ("to be").

In the lexicon, lexical units such as *estar_VB_04* are described; this fourth entry for *estar* corresponds to a verb that has two arguments, the second being an adverb or a prepositional group. *estar_VB_01* is the simple copula, *estar_VB_02* is the existential *be*, which has only one argument, and *estar_VB_03* is the auxiliary. Each lexical unit contained in the concepticon is a key in the lexicon. The lexicon has been crafted manually for the experiments in this paper, but we are developing an automatic conversion of AnCora-Verb (Aparicio et al., 2008) to obtain a large scale resource. Finally, in order to store the surface forms of the inflected words, we crafted a very small morphological dictionary of about 450 entries to cover the needed forms in the experiments.

4.4 Obtaining Spanish property values

The DBpedia project uses the Resource Description Framework (RDF) as a data model for representing and publishing on the Web structured information that has been extracted from Wikipedia. Each DBpedia entity (*resource*) is directly tied to a Wikipedia article and denoted using a de-referenceable URI or IRI. Until DBpedia release 3.6, data were extracted from non-English Wikipedia pages only if an equivalent English page existed, in order to ensure that each entity is uniquely identified by a single de-referenceable URI of the form `http://dbpedia.org/resource/Name` (e.g., `http://dbpedia.org/page/Switzerland`), where *Name* is derived from the URL of the source (English) Wikipedia article. As of DBpedia release 3.7, localized datasets are provided that contain data from all Wikipedia pages in a specific language, using IRIs and language-specific namespaces of the form `http://xx.dbpedia.org/resource/Name`, where 'xx' is the Wikipedia language code and 'Name' is now derived from the respective language-specific Wikipedia URL, e.g., `http://es.dbpedia.org/page/Suiza`; inter-language links from the different Wikipedia editions are also extracted and the `owl:sameAs` property is used to link the localized DBpedia IRI to its equivalent in English DBpedia edition URI.

Thus, whenever an inter-language link between a non-English Wikipedia page and its English equivalent exists, by querying the `owl:sameAs` property links of the English DBpedia entity and filtering them using the language code, respective language-specific names can be obtained. However, not every English Wikipedia page has an equivalent page in every non-English Wikipedia edition; moreover, even if an equivalent non-English page exists, the respective `owl:sameAs` link does not necessarily pertain to the English DBpedia entity at hand (as for example, in the case of the Spanish entity `http://es.dbpedia.org/page/Galleta` that can be accessed only when starting from the English resource `http://es.dbpedia.org/page/Biscuit`, but not from `http://es.dbpedia.org/page/Cookie`). Further complications may still arise, as sometimes the obtained language-specific name corresponds to the most rigorously rather than commonly used name, which, in the context of NLG, can affect the fluency of the resulting verbalization; for example, starting from the English entity *Chicken*, the Spanish value is *Gallus gallus domesticus*, instead of *Gallo*. Moreover, sometimes datatype values (i.e., raw data) rather than entities are used as object values (e.g., `Bakewell_pudding` || *ingredientName* || ``Ground almond, jam, butter, eggs''.

4.5 Improving language-independent rules

FORGe received good evaluation marks at the WebNLG challenge, especially in the human assessments, according to which it was close to the quality of human-written text. However, after an error analysis of FORGe's outputs, we found a series of general problems impairing the quality of the generated texts in terms of contents and grammaticality. In particular: (i) some properties were not verbalized due to the failure to produce relative clauses in some specific cases; (ii) the aggregations were at times excessive, erroneously merging verbs with different tenses (e.g. *X created Y, which was created by Z*, instead of *X and Z created Y*), failing to merge (e.g. *X is the headquarters of Y. Z*

is the headquarters of Y), or leading to an ungrammatical outcome, with for instance the presence of several *also*; (iii) the construction of some relative clauses were faulty, as e.g. *X can a variation of which be Y*, instead of *X, which can be a variation of Y*; (iv) the referring expression module was applying excessively, resulting in ambiguous pronouns, and sometimes incorrectly pronominalizing non-human entities with *he*; (v) some agreements were not solved (e.g. *the main ingredient are*); (vi) some determiners were erroneously introduced, and some others not in the correct form (*a* instead of *an*). And for English in particular, (vii) some templates were mixed up (e.g. runway name with runway number), and some were incorrect, with present instead of past tense.

Many occurrences of these issues were fixed in the grammars, by modifying and adding rules, and some new features were added, as for instance, new aggregation and pronominalization types in order to improve the fluency of the outputs, and new rules to cover more cases of embedded clauses generation. For developing the grammars, we used the 6 and 7 triple inputs from the WebNLG training data, and the whole development set. A qualitative evaluation of the new outputs is provided in Section 6.

As a result, the extended version of the DBpedia generator comprises 971 active rules. 73% of the rules (702) are language-independent, 19% are for English, and 8% for Spanish.[9] For instance, all (82/82) of the aggregation rules and most (365/395) of the sentence structuring rules, which map PredArg graph onto Deep-Syntactic graphs, apply for both languages. When getting closer to the surface, the rules are less language-independent, representing about half of the DSynt-SSynt rules (108/239) and of the linearization and agreement resolution rules (66/129).

5 Evaluation

In this section, we detail how we built a new dataset for evaluating the outputs, and describe the results of the automatic evaluations.

5.1 Selection of triples for evaluation

For evaluation purposes, we compiled a benchmark dataset of 200 inputs, i.e., sets of DBpedia triples, with sizes ranging from 1 to 7 triples, using as reference pool the WebNLG challenge test set. The reason for using as reference basis the WebNLG challenge dataset is that it is the most recent and comprehensive dataset with respect to text generation from RDF data that has been specifically designed to promote data and text variety (Perez-Beltrachini et al., 2016). Moreover, it allows the direct comparison with the generators that participated in the challenge. In order to ensure future comparisons with machine learning-based systems in terms of their best obtained performance, only the seen categories subset of the original test set has been considered, i.e., only inputs with entities that belonged to DBpedia categories that were contained in the training data.

The compilation methodology for our benchmark dataset implements a twofold goal. On one hand, we want to ensure that all properties appearing in the seen categories subset are included. On the other hand, and unlike the WebNLG human evaluation test set, we aim towards a more balanced number of inputs of different sizes. In practice, since the inputs of size 6 and 7 in the original seen categories subset of the WebNLG test set are 24 and 21 respectively, we chose to include them all in the benchmark; 31 inputs for each of the remaining input sizes were subsequently added.

5.2 Reference sentences

The English reference texts are taken from the WebNLG dataset, for which there could be more than one reference per triple set. For Spanish, one single reference text was produced for each triple set, with natural and grammatical constructions containing all and only the entities and relations in the triples. The reference texts were written by one of the authors, a native Spanish speaker, having at hand the English references from the WebNLG challenge to serve as a potential model.

5.3 Automatic evaluation

The predicted outputs in English and Spanish were compared to the reference sentences in the corresponding language; three metrics were used: BLEU (Papineni et al., 2002), which matches exact words, METEOR (Banerjee and Lavie, 2005), which matches also synonyms, and TER (Snover et al., 2006), which reflects the amount of edits needed to transform the predicted output into the reference output. Table 1 shows the results of the automatic evaluation on the English and Spanish

[9] Note that we exclude from the count all rules than simply transfer individual attributes at each level, which amount to about 250. There are more English-specific rules simply because the coverage of the English generator is higher.

extensions proposed in this paper using for each input its corresponding reference text(s). The first two rows show that in terms of automatic metrics, the extended FORGe and the 2017 FORGe have almost exactly the same scores on the English data (which are also very close to the WebNLG scores: 40.88, 0.40, 0.55). In other words, the quality improvements in English are not reflected by these metrics. To compare English and Spanish results, we calculated the scores using one sentence as reference (only one reference per text is available in Spanish). The English scores drop (third row) due to the way the scores are calculated by the individual metrics.[10] In the last row of the table, the scores of the Spanish generator look contradictory: the BLEU is 10 points below the English BLEU with the same number of reference (1), but METEOR is 8 points above, that is, the predicted outputs do not match the exact word forms, but they do match similar words. One reason for the low BLEU score could be the higher morphological variation in Spanish. However, the METEOR score is surprisingly high, actually even higher than the highest METEOR score at WebNLG, obtained by ADAPT and calculated with multiple references (0.44).

Reference set	BLEU	METEOR	TER
EN ($All_{FORGe-2017}$)	39.87	0.40	0.58
EN ($All_{FORGe-Ext}$)	39.33	0.40	0.58
EN ($1_{FORGe-Ext}$)	29.18	0.38	0.65
ES ($1_{FORGe-Ext}$)	18.68	0.46	0.77

Table 1: English and Spanish scores according to BLEU, METEOR and TER, with 1 and All references on the 200-triples test set.

6 Qualitative analysis of the results

In the 200 outputs of the 2017 generator, 275 errors were detected, compared to 166 in the current one in English (170 in Spanish), and 26.5% of the texts were error-free, as opposed to 43.5% now (45.5% in Spanish). In this section, we report on the examination of both English and Spanish outputs, in order to identify the main issues of the grammars in both languages.[11]

6.1 English

The qualitative analysis of the generated English texts showed that the resulting texts are of a higher grammaticality and fluency than the 2017 ones. Below, we discuss the observed remaining errors and their respective causes.

Determiners: Definite determiners are missed with the property `language`, when referring to the language of a written work. The reason of this error lies in the discrepancy between the respective PredArg template that was defined based on the premise that the object value of this property is a language name (i.e., English, Italian), hence not admitting a determiner, and the form of the DBpedia language entities that in practice concatenate the language name with the word *language* (cf., *English_language*); this type of error is the most frequent, being found about 65 times in the test set and representing about 40% of the total amount of errors (166). This underlies the need for further normalization of the DBpedia property values, so that during the PredArg templates instantiation, consistent linguistic features will be ensured for argument values of the same type.

Tense: Errors are observed with respect to the verb tense selection (6% of the errors). More specifically, in some cases the present tense is used instead of the past, as, e.g., in *Alan Shepard, who graduated from NWC in 1957 with a M.A., is deceased. [...] He is a test pilot.* This is a direct consequence of the fact that in the current implementation, tense selection does not take into account the temporal context as defined by the rest of the input triples.

Aggregations: Another type of error relates to the generation of unintuitive, yet still grammatical, constructs when aggregating the contents of more than one triple when certain properties are involved (11% of the errors). More specifically, when the property `occupation` is selected to be expressed as a relative clause, it fails to append the occupation information to the referring entity as shown in *Alan Bean, born in wheeler (Texas) on March 15, 1932, is from the United States (test pilot)*; a similar behaviour has been observed with the property `category`. This is a result of the current implementation of aggregation that takes place in a single step and tries to avoid orphan clauses by attaching them to the closest reference head; introducing iterative aggregation steps and incorporating semantic coherence information would mitigate such effects.

A related issue is, for instance, the way location information is verbalized in the presence of multi-

[10] BLEU matches n-grams in all candidate references, and METEOR and TER consider the best scoring reference.

[11] Outputs are available as supplementary material below.

ple subdivision references (15% of the errors), as, for example, in *the Acharya Institute of Technology is in Bangalore, Karnataka and India*, where the three involved location-denoting properties, namely `city`, `state` and `country` have been aggregated in a semantics-agnostic manner. Navigating DBpedia and obtaining information about their interrelations would enable more fluent verbalizations. Fluency and meaning accuracy are also impacted when the input triples capture in practice n-ary relations. This is the case with the `leader` and `leaderTitle` properties, which in the absence of any semantic preprocessing before the instantiation of the PredArg templates, result in verbalizations such as *the leaders of Romania are the prime minister of Romania and Klaus Iohannis*, which does not communicate the fact that Klaus Iohannis is the prime minister.

Subject/Object values: Last, a number of disfluent verbalizations is the direct result of idiosyncrasies in the involved DBpedia properties and/or the respective subject and object values (4% of the errors). There are properties that although meant to capture different types of information are not used consistently, thus impacting the resulting verbalizations; the properties `mainIngredient(s)` and `ingredient(s)` are such an example, e.g. in an input about the dish *Ayam Penyet*, which is described as having as main ingredient the fried chicken and as a further ingredient chicken. Some minor errors such as unnatural word ordering (11%) or lexicalizations (8%) were also detected.

6.2 Spanish

The aforementioned errors listed for English are mostly independent of the language and thus also apply to Spanish, except from the first aggregation error, which does not appear due to a difference in the templates. The determiner error represents 30% of the total number of detected errors (51/170), the location aggregation 12%, the values and word choices 7%, the ordering 6%, the verbal tense 5%. However, despite its overall good quality, Spanish has some additional specific issues.

English words: There are some not-translated nouns (*52 minutes*) or phrases (*está dedicado a Ottoman army soldiers killed in the battle of Baku*), which in addition of not being understandable, may produce subsequent morphological errors (21% of the errors).

Morphology: Morphological errors, mainly gender (invisible in English) and number disagreements, are found in the Spanish texts (5% of the errors). For example, in *Dianne Feinstein es un senador de california*, (lit. 'Dianne Feinstein is a$_{MASC}$ senator$_{MASC}$ of California'), both *a* and *senator* should be feminine, but there is no information that D. Feinstein is a woman in the input.

Complex relative clauses: The main syntactic error is related to the genitive relatives with *cuyo* ('of which'), in particular when the antecedent is a location (5% of the errors). For example, in the sentence *Alba Iulia , en el cual está el 1 Decembrie 1918 University*, lit. 'Alba Iulia, in the which is the 1 Decembrie 1918 University', the proper pronoun should be *donde* 'where' instead of *en el cual*. Even when gramatically correct, sentences with these relative clauses tend to lack naturalness.

Other series of errors that produce sub-optimal Spanish constructions include: occasional choice of a relative clause instead of a past participle modifier, and various other constructions that lack naturalness (10% of the errors).

7 Conclusions and future work

This paper reports on the extension of the FORGe system for verbalizing DBpedia triples, which results in a better quality of English texts, and the adaption of FORGe to Spanish. The qualitative evaluation of both English and Spanish texts showed that overall, the grammaticality and fluency of the resulting verbalizations was high, but could be further improved, in particular by getting more information about the subject and object entities. The next step is to run a large-scale human assessment of the outputs in terms of quality of language and contents. Furthermore, the DBpedia cross-language overlap is not sufficiently high to obtain property values in languages other English by using only inter-language links; in our evaluation, it approximated 55%, but this percentage can vary, depending on how well-known the referred entities are, thus requiring complementary investigations. Another objective is to port FORGe to other languages.

Acknowledgements.

This work has been partly supported by the European Commission (H2020 Programme) under the contract numbers 700475-IA, 700024-RIA, 779962-RIA, 786731-RIA and 825079-ICT-STARTS. We thank Anastasia Shimorina and the reviewers for their valuable help and feedback.

References

Ion Androutsopoulos, Gerasimos Lampouras, and Dimitrios Galanis. 2013. Generating natural language descriptions from owl ontologies: the naturalowl system. *Journal of Artificial Intelligence Research*, 48:671–715.

Juan Aparicio, Mariona Taulé, and M Antònia Martí. 2008. Ancora-verb: Two large-scale lexicons for catalan and spanish. In *Proceedings of the XIII Euralex International Congress*. Institut Universitari de Lingüística Aplicada, UPF.

Olga Babko-Malaya. 2005. *Propbank Annotation Guidelines*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Anja Belz, Mike White, Dominic Espinosa, Eric Kow, Deirdre Hogan, and Amanda Stent. 2011. The first Surface Realisation Shared Task: Overview and evaluation results. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation (ENLG)*, pages 217–226, Nancy, France.

Bernd Bohnet and Leo Wanner. 2010. Open soucre graph transducer interpreter and grammar development environment. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

M. Bollmann. 2011. Adapting SimpleNLG to German. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG 2011)*, pages 133–138.

Nadjet Bouayad-Agha, Gerard Casamayor, and Leo Wanner. 2014. Natural language generation in the context of the semantic web. *Semantic Web*, 5(6):493–513.

R. De Oliveira and S. Sripada. 2014. Adapting SimpleNLG for Brazilian Portuguese realisation. In *Proceedings of the 8th International Natural Language Generation Conference*, pages 93–94.

Henry Elder, Sebastian Gehrmann, Alexander OConnor, and Qun Liu. 2018. E2e nlg challenge submission: Towards controllable generation of diverse natural language. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 457–462.

Claire Gardent and Aldo Gangemi. 2016. Proceedings of the 2nd international workshop on natural language generation and the semantic web (webnlg 2016). In *Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017a. Creating training corpora for micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017b. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133.

A. Gatt and E. Reiter. 2009. SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 90–93.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

R. de Jong and M. Theune. 2018. Going Dutch: Creating SimpleNLG-NL. In *Proceedings of the 11th International Natural Language Generation Conference*, pages 73–78.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, pages 1989–1993, Las Palmas, Canary Islands, Spain.

François Lareau, Florie Lambrey, Ieva Dubinskaite, Daniel Galarreta-Piquette, and Maryam Nejat. 2018. Gendr: A generic deep realizer with complex lexicalization. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 3018–3025, Miyazaki, Japan.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

A. Mazzei, C. Battaglino, and C. Bosco. 2016. SimpleNLG-IT: Adapting SimpleNLG to Italian. In *Proceedings of the 9th International Natural Language Generation Conference*, pages 184–192.

Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press, Albany.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank Project: An interim report. In *Proceedings of the Workshop on Frontiers in Corpus Annotation, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pages 24–31, Boston, MA, USA.

Simon Mille, Alicia Burga, and Leo Wanner. 2013. AnCora-UPF: A multi-level annotation of Spanish. In *Proceedings of the 2nd International Conference on Dependency Linguistics (DepLing)*, pages 217–226, Prague, Czech Republic.

Simon Mille and Stamatia Dasiopoulou. 2017. *FORGe at WebNLG 2017*. Technical report, Dept. of Engineering and Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain.

Simon Mille and Leo Wanner. 2015. Towards large-coverage detailed lexical resources for data-to-text generation. In *Proceedings of the First International Workshop on Data-to-text Generation*, Edinburgh, Scotland.

Neha Nayak, Dilek Hakkani-Tür, Marilyn A Walker, and Larry P Heck. 2017. To plan or not to plan? discourse planning in slot-value informed sequence to sequence models for language generation. In *Proceedings of INTERSPEECH*, pages 3339–3343, Stockholm, Sweden.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Laura Perez-Beltrachini, Rania Sayed, and Claire Gardent. 2016. Building rdf content for data-to-text generation. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Alejandro Ramos Soto, Julio Janeiro Gallardo, and Alberto Bugarín Diz. 2017. Adapting simplenlg to spanish. In *Proceedings of the 10th International Natural Language Generation Conference (INLG)*, pages 144–148.

Leo Wanner, Bernd Bohnet, Nadjet Bouayad-Agha, Francois Lareau, and Daniel Nicklaß. 2010. MARQUIS: Generation of user-tailored multilingual air quality bulletins. *Applied Artificial Intelligence*, 24(10):914–952.

A Supplementary Material

Sample sentences with no detected problems.

ES: Antioch (California), cuya población total es 102372, tiene una diferencia horaria UTC de -7. Su código de área es 925. La superficie total de Antioch (California) es 753 kilómetros cuadrados.

EN$_{Ext}$: Antioch, California, the total population of which is 102372, has a UTC offset of -7. Its area code is 925. The total area of Antioch, California is 753 square kilometers.

EN$_{2017}$: Antioch, California, the total population of which is 102372, has a UTC offset of -7. The area code of Antioch, California is 925. The total area of Antioch, California is 753 square kilometers.

Sample sentences with a missing translation and unnatural word ordering (ES).

ES: 1634: The Ram Rebellion se publica en Estados Unidos. Barack Obama es el líder de Estados Unidos, en el cual viven americans. La capital de Estados Unidos, un grupo étnico del cual es afroamericanos, es Washington D.C..

EN$_{Ext}$: 1634: The Ram Rebellion is published in the United States. Barack Obama is the leader of the United States, in which Americans live. the capital of the United States is Washington (D.C.). An ethnic group of the United States are African Americans.

EN$_{2017}$: 1634: The Ram Rebellion is published in the United States. Barack Obama is the leader of the United States, Americans live in which. The capital of the United States is Washington (D.C.). An ethnic group of the United States are African Americans.

Sample sentences with a missing translation (ES) and determiner error (EN, ES).

ES: Asilomar Conference Grounds, que fue incorporado al National Register of Historic Places el 27 de febrero de 1987, está en Pacific Grove. Su número de referencia en registro nacional de sitios históricos es 87000823. Asilomar conference grounds se construyó en 1913.

EN$_{Ext}$: Asilomar Conference Grounds, which was added to the National Register of Historic Places on February 27 (1987), is in Pacific Grove (California). The reference number in the National Register of Historic places of Asilomar Conference Grounds, built in 1913, is 87000823.

EN$_{2017}$: Asilomar Conference Grounds, the reference number in the National Register of Historic Places of which is 87000823, is in Pacific Grove (California) in added to the National Register of Historic Places on February 27 (1987). It was built in 1913.