Multilingual Generation of Noun Valency Patterns for Extracting Syntactic-Semantical Knowledge from Corpora (MultiGenera)

María José Domínguez Vázquez[1], Carlos Valcárcel Riveiro[2], David Lindemann[3]
[1]Universidade de Santiago de Compostela, [2]Universidade de Vigo, [3]Universität Hildesheim
E-mail: majo.dominguez@usc.es, carlos.valcarcel@uvigo.es, david.lindemann@uni-hildesheim.de

Abstract

The aim of the project is to develop a prototype for a generator of argument structure or valency realizations in terms of syntagmatic and paradigmatic combinations of Spanish, German and French nouns. The two main applications of the tool prototype we are aiming to develop, are (1) the generation of noun phrases as argument structure realizations that follow patterns related to semantic features, for the creation of corpus and web query strings; and (2) the knowledge-based generation of simple and complex noun phrases that are acceptable in a coherent sentence context. An essential step in developing these applications is the systematic description of the valency-related syntagmatic and paradigmatic properties of argument combinations. To this end, we have devised a methodology based on bidirectional mutual enrichment (bottom-down and bottom-up). With the aim of generating argument surface realizations, we will mainly use lexical knowledge represented in wordnets for Spanish, German and French for the semantic annotation of lexical prototypes and their subsequent paradigmatic expansion.

Keywords: noun valency, argument structure, combinatory patterns, corpus lexicography, ontologies, wordnet, natural language generation

1 Introduction: From PORTLEX to MultiGenera

The work done for the PORTLEX dictionary[1] (*cf.* Domínguez & Valcárcel, in press) has highlighted the limitations of corpus-driven methods in achieving the objective of a lexicographical project, at least from a dependency-valency perspective: to compile all acceptable types of constructions (Mel'čuk 2013). Corpora do not contain examples of all the realizations of the different actants or of their combinations. On the other hand, the examples found in corpora sometimes do not meet the intelligibility or conciseness requirements of a dictionary. Furthermore, the available corpora of a significant size are not semantically annotated, so that it is not possible to apply semantic filters when searching for syntactic patterns. For example, we may search for *muerte de* [NP] *por* [NP] ('death of [NP] by [NP]'), but we may not filter this search by certain semantic values for each of the [NP] slots, like, for example, *death of* [noun: +living being] *by* [noun: +disease], i.e. items the hyponyms of which typically would fit into that slot as argument. On the other hand, there will be examples found in corpora that formally do fit into the queried valency pattern, but that are not argument realizations. For example, the noun+adjective combination *control paterno* ('*parental control*') does constitute a realization of verb and 'agent', while *control férreo* ('*rigid control*'), also noun+adjective, does not, since the adjective here is a mere attribute to the noun (cf. Domínguez 2011). A further difficulty is linked to the fact that certain combinations of actants cannot be found in most corpora. For example,

1 PORTLEX (*Diccionario multilingüe de la frase nominal* / *Multilingual dictionary of the noun phrase*), accessible at http://portlex.es.

in French *La saveur chocolat de vos gâteaux* 'The chocolate flavor of your cakes', which is documented in corpora, versus *La saveur citron de vos gâteaux* 'The lemon flavor of your cakes', which is not documented, despite being semantically and syntactically similar to the previous one and no less acceptable for a French speaker.

The multilingual tool prototype MultiGenera is designed according to the principles of Valency Grammar and Lexicography (Engel 1995), and following related work on wordnets, i.e. concept-based lexical resources represented as ontologies. This will allow an analysis of different lexical-semantic domains in Spanish, German, and French from an onomasiological-conceptual point of view. The development of such a tool thus involves a contrastive approach to nouns as lexical items that belong to certain lexical-semantic domains. In other words, we propose developing a tool for processing syntagmatic and paradigmatic combinations driven by data extracted from corpora and wordnets.

MultiGenera will apply semantic filters to the results of searches in semantically not annotated corpora, using wordnet ontologies (see the references in Section 2.2) for the different languages under analysis. Existing and newly developed tools will be combined in order to obtain detailed syntactic-semantic information. The tool therefore randomly generates realizations of an argument structure (i.e. in Spanish *el olor a* [noun: +animate/+material] *de* [noun: +material][2]), by establishing a connection between semantic features ([+material], [+animated], etc.), on one side, and concepts belonging to wordnet ontology categories or that are hyponyms to wordnet concepts that represent the desired semantic feature, on the other: *el olor a humedad de su habitación* ('the musty smell of his room'), *el olor a hombre de la chaqueta* 'the jacket's smell of man', *el olor a caballo del establo* ('the smell of horse in the stable'), etc. In order to filter out unacceptable data, generated realizations are then intersected with search results in corpora. However, our methodology also allows us to detect, through expert verification in each language, realizations and combinations that do not appear in corpora or on the web (for different reasons), but that are possible and acceptable. This would enable us to in some way overcome the limitation of corpora as an attestation method, since they only show a limited part of the combination possibilities of a language.

In many cases, it will also generate unacceptable sequences, but this is also a particularly interesting aspect of our research:

All in all we have received several hundred paraphrase clusters on the computer. Some of them contained serious mistakes, and it is precisely those clusters that were of exceptional interest to us. A linguistic processor incorporating a serious linguistic theory becomes, by the very nature of things, a gigantic testing ground for this theory. The computer makes mistakes unimaginable for a human. The analysis of such mistakes can be extremely revealing in the sense that it is a shortcut to correcting lexicographic and grammatical descriptions of which its linguistic software is composed. Moreover, very often experimenting with formal models of language on the computer results in genuine linguistic discoveries (Apresjan et al. 2003, p.11).

2 Methodology and Workflow

The design and development of MultiGenera follows five core working steps, the description of argument structure realization patterns (2.1.), expansion of lexical prototypes (2.2.), the generation of argument structure realizations (2.3.), argument combinations within noun phrases (2.4.), and context generation (2.5.).

2 In English, 'the [adjective: +animate, +material] smell of [noun: + material]' or 'the smell of [noun: +animate, +material] from/in [noun: +material]'. For example: *el olor a tabaco de la habitación* 'the smoke smell of the room' or 'the smell of smoke from/in the room'.

2.1 Description of Realization Patterns for Argument Structures

The starting point for this description is the set of valency patterns provided in PORTLEX for the ten selected nouns. PORTLEX valency schemata contain patterns for argument structure realization in syntactic slots, in several possible combinations, together with examples for the filling of the slots with lexical items. The following table shows some patterns provided by PORTLEX for the German noun *Geruch* ('smell') together with surface realization examples; note that the syntactic slots of arguments are annotated with semantic categories that will later allow the extraction of lexical data using lexical-semantic relations encoded in wordnet ontologies.

Table 1: Argument structures and semantic features for the German noun *Geruch*.

Det.	***{Adjective}***[3]	***Noun Phrase Head***	***{Genitive Det.}***	***Noun A1***[4] ***[Material]***
Der	angenehme	Geruch	der	Blumen
The	*pleasant*	*smell*	*of the*	*flowers*
Det.	***{Adjective}***	***Noun Phrase Head***	***von (+ {Det.})***	***Noun A1 [Material]***
Der	intensive	Geruch	von diesen	Männern
The	*intense*	*smell*	*of these*	*men*
Det.	***{Adjective} A1***	***Noun Phrase Head***	***nach(+ {Det.})***	***Noun A2***[5] ***[Material]***
Der	menschliche	Geruch	nach	Schweiß
The	*human*	*smell*	*of*	*sweat*
Det.	***{Adjective}***	***Adj. A1 [Animate]***	***Noun Phrase Head***	
Der	intensive	männliche	Geruch	
The	*intense*	*male*	*smell*	
Det.	***{Adjective}***	***Noun A1 [Material]***	***Noun Phrase Head***	
Der	stechende	Schweiß	-geruch	
The	*pungent*	*sweat*	*smell*	

2.2 Expansion of Lexical Prototypes

The slot-filling lexical items listed in the PORTLEX schemes will be expanded by a list of lexical prototypes, i.e. frequent nouns or adjectives that belong to a semantic category that corresponds to the specified semantic role of an argument. In other words, for each argument-role-slot a general list of prototypical lexical items will be obtained, as shown in Table 2 for the Spanish argument structure Det. + *olor a* + common noun (*aquel olor a tabaco*, 'that smell of tobacco').[6] These lexical items will be collected by queries in *Sketch Engine*,[7] which provides large corpora for the three languages, together with frequency data.

Slot-filling prototypes found in corpora for every argument are associated with a semantic category following an ontology of semantic features, which has been specifically designed for MultiGenera. Four different levels are differentiated, ranging from the most general to the most specific.[8] Table 3

3 Curly brackets mean that an item does not appear necessarily according to the valency pattern.

4 A1' refers to the argument with the meaning 'someone or something, that has something'. In this case: The flowers have a pleasant smell.

5 'A2' refers to the argument with the meaning 'something belongs to a class or type'. In this case: The smell is of sweat.

6 The list is the result of an examplary query to *eseuTenTen11* corpus using *Sketch Engine*. Results for French were obtained from *frTenTen12* corpus.

7 See http://the.sketchengine.co.uk/.

8 The first two levels cover the following main semantic features: [Material]: [substance] and [objects]; [Animated]: [human], [animal], [fungus] and [plants]; [Situation]: [static situations], [processes], [locations]; [Intellectual concepts]. The third and the fourth levels are more specific but they are not always applicable (see Table 3).

Table 2: Example of ranking ten lexical prototypes for the Spanish argument structure *olor a* + common noun.

Prototypes (nouns)	Corpus counts
tabaco ('tobacco')	235
incienso ('incense')	177
pólvora ('gunpowder')	155
humo ('smoke')	153
humedad ('humidity')	142
gasolina ('petrol')	132
azahar ('orange blossom')	92
sudor ('sweat')	90
azufre ('sulfur')	87
naftalina ('naphthalene')	79

shows an example of the semantic annotation carried out for MultiGenera concerning the exemplary lexical prototypes displayed in Table 2:

Table 3: Example of semantic annotation of lexical prototypes for the Spanish argument structure *olor a* + common noun.

Lexical prototypes	1st Order	2nd Order	3rd Order	4th Order
tabaco ('tobacco')	Material	Substance	Solid	Smoke
incienso ('incense')	Material	Substance	Solid	Chemical
pólvora ('gunpowder')	Material	Substance	Solid	Chemical
humo ('smoke')	Material	Substance	Gas	Smoke
humedad ('humidity')	Situation	State	Property	
gasolina ('petrol')	Material	Substance	Liquid	Fuel
azahar ('orange blossom')	Animate	Plant	Flower	
sudor ('sweat')	Material	Substance	Liquid	Excrement
azufre ('sulfur')	Material	Substance	Liquid	Excrement
naftalina ('naphthalene')	Material	Substance	Solid	Chemical

As a result, a conceptual map of the acceptable values for an argument-role slot can be visualized, showing also contrasts from language to language. For the argument A2 of the French noun *odeur* 'smell', for example, there are essentially three main semantic classes of lexical prototypes: [+Material, +Substance] (*sueur* 'sweat', *tabac* 'tobacco', *poudre* 'gunpowder'), [+Animate, +Plant] (*fleur* 'flower', *jasmin* 'jasmine'), and [+Material, +Object] (*pain* 'bread', *crêpe* 'crepe'). By manually associating the MultiGenera semantic category descriptors with wordnet synsets, we can validate the semantic relation (hyponymy, meronymy) between the category descriptor and the lexical prototypes, i.e. their semantic annotation, including disambiguation of word senses.

Beyond its usefulness for describing the semantic features of a nominal argument, this annotation process also makes it easier to expand the prototype lists using item relations encoded in wordnet. For each semantic class of prototypes we search for correspondences with categories or subcategories in wordnet ontologies. Thus, for example, for the group of lexical prototypes [+Material, +Object, +Food] of the argument A2 of the Spanish noun *olor* or its French equivalent *odeur*, it has been possible to establish a connection with the subcategory [+Food] of the TOP ontology. For the task of exploring the different semantic relationships with which wordnets operate, a specific API for database queries has been developed for each language of the project. Connections are not only established between groups of lexical prototypes and categories of the TOP (Rodríguez et al. 1998) and

SUMO ontologies (Niles & Pease 2003), but also with WordNet Domains (Gonzalez, Rigau & Castillo 2012)) and epinonyms (Guinovart & Solla 2018), and using hyponymy or meronymy relations encoded in wordnets for the three languages.[9]

When a link to wordnet items is set, all the existing items in the pertaining ontological category are extracted to form a set of candidates as argument slot fillers. In the case of the argument A2 for *olor* and *odeur*, all items belonging to the subcategory [+Food] of the TOP ontology will be considered. All candidate lists will be validated in two steps: an automated and a manual one. In automated data validation, an intersection is made between the list of lexical items in a subcategory or hyponym cluster in Wordnet and the lexical items collected from corpora for a specific syntactic slot. All items present in the corpus are automatically validated and the remainder are manually validated by experts in each of the languages of the project. Thus, following the previous example, items such as *liqueur* 'liqueur' (present in the corpus) and *hachis* 'hashish' (not present) would remain in the set, while *vitamine B2* or *vendange* 'grape harvest' would be excluded. The result is an expansion of the initial group of lexical prototypes that can go beyond the limitations of corpora.

2.3 Generation and Assessment of Single-Argument Surface Realizations

The automatic generation of the argument structures of the ten selected nouns to develop the prototype implies, once again, the joint use of several tools. In this case, in addition to wordnets for the extraction of the lexical knowledge, *Freeling*[10] will be used for the introduction of morphological data (gender, number and, for German, also the case). Noun phrases in Spanish, French and German are subject to rules of agreement between the determiner and head. Moreover, these languages often present contractions of prepositions and the articles that follow them. For example, for the argument A1 of *olor* we have: *el olor del pan fresco* ('the smell of the fresh bread').

For the automatic generation of the argument structures, we use own python scripts and our own API for accessing wordnet and semantic ontologies. The lists of candidate lexical items to fill in each argument slot will allow users of our tool to choose those they prefer to generate simple noun phrases.

2.4 Argument Combinations within Noun Phrases

We also aim to generate argument combinations within a noun phrase, as, for example, an argument structure with two semantic roles realized in prepositional phrases (e.g., in Spanish *el olor a sudor de tu ropa* 'the smell of sweat from your clothes'). The aim is to assess the combinatorial compatibility of the paradigmatic sets defined before (see Section 2.3). In this way, the candidate lists for each argument will be combined with those of the other noun arguments in different positions within the noun phrase. This raises the issue of semantic constraints governing combinations of arguments. Separately acceptable semantic categories for filling in different argument slots can present problems when combined in the same nominal phrase. Thus, for the noun *olor*, 'smell', it is usually not possible to combine arguments A1 and A2 belonging respectively to the categories [+Animal] and [+Food]: (**el olor a tomate de los ratones* 'the tomato smell from mice', **el olor cárnico del cocodrilo* 'the meaty smell of the crocodile', **el olor del caballo a tortilla* 'the horse's smell of omelet').

Although the combinations already registered in PORTLEX will be used as a starting point, we suggest that the acceptability of argument structure realizations not attested in corpora can be validated by this method. As many acceptable argument combinations are not included in corpora, the

9 At present, we use MCR 3.0 (Gonzalez, Laparra & Rigau), available at http://adimen.si.ehu.es/web/MCR, and the Extended Open Multilingual Wordnet (Bond & Foster 2013), http://compling.hss.ntu.edu.sg/omw/summx.html.

10 See http://nlp.lsi.upc.edu/freeling/.

assessment of the generated argument combinations will have to be carried out manually by members of the project team. At the end of this phase, an exhaustive list of the acceptable combinations of arguments will be obtained for each noun, as well as valuable information on grammatical constraints.

2.5 Context generation

The main objective at this fifth stage is to create contexts of whole sentences for the generated nominal phrases. The results of a preliminary study on the noun *muerte* 'death' (Valcárcel & Domínguez 2016), where the acceptability of the generated nominal phrases was assessed by anonymous participants, suggests that the assessment of semantic acceptability of automatically generated nominal phrases could be improved by providing a whole-sentence context. This conclusion led us to MultiComb, a parallel project to MultiGenera but longer in duration. MultiComb, which is funded by the Spanish Ministry of Economy, Industry and Competitiveness, is focused on generating more acceptable and familiar output for a human speaker. It is necessary to distinguish here the generation of context at the phrase level, on one hand, and at sentence level, on the other.

2.5.1 Context Generation at the Phrase Level

For context generation at the phrase level, we add adjective attributes to the ten nouns selected for developing the prototype within MultiGenera (i.e. in Spanish *un fuerte olor a tabaco* ('a strong smell of tobacco'), *aquel agradable olor a madera de su habitación* ('that pleasant smell of wood in his room'). For a formal modelling and machine-readable annotation of these structures, a selection of basic lexical functions (LF) related to qualifiers is carried out, following the proposal of Mel'čuk (2013, 2015). In similarity to the working steps described above for the extraction of slot-filling lexical items, a selection of lexical prototypes according to frequency in corpora will allow the definition of paradigmatic sets for each LF. The lexical items to represent the paradigmatic restrictions will be collected from corpora and collocation dictionaries for the three languages involved. Obviously, these paradigmatic sets associated with LF will depend not only on each noun, but also on the specific lexical restrictions of each of the three languages. For example, in the case of the Spanish noun *olor* 'smell', we would obtain the following prototype lists for the selected LF:

- Magn (*olor*) = *fuerte ('strong'), intenso ('inten*se'), *penetrante* ('pungent, penetrating')
- AntiBon (*olor*) = *malo* ('bad'), *desagradable* ('unpleasant'), *nauseabundo* ('nauseating'), *rancio*, ('rancid'), *insoportable* ('unbearable'), *asqueroso* ('nasty'), *fétido* ('foul')
- Bon (*olor*) = *agradable* ('pleasant'), *fresco* ('fresh'), *dulce* ('sweet')
- Ver (*olor*) = *característico* ('characteristic'), ge*nuino* ('genuine'), *verdadero* ('real')

This allows us to randomly program the appearance of adjectives linked to a noun by an LF, and obtain a more varied and human-like output. Again, the issue of semantic constraints arises here. For example, some semantic categories of the argument A1 of the Spanish *olor* ('smell') such as [+Flower] imply Bon adjectives while others such as [+Excrement] demand AntiBon qualifiers (see Table 4). Thus, it is at least striking for a speaker to hear or read a nominal phrase such as *el agradable* (Bon) *olor de las cloacas* [+Place, +Building, +Excrement] ('the pleasant smell of the sewers').

2.5.2 Context Generation at the Sentence Level

In this stage, the previously generated noun phrases (Det + noun + arguments) will fill in the valency slots of a verb. These sentence contexts will be limited to four basic syntactic structures: [Subject (NP) + Verb: *el olor a tabaco de la casa se disipó*, 'the tobacco smell in the house faded away'], [Subject (NP) + Copula + Attribute: *el olor a tabaco de la casa resultaba insoportable*, 'The tobacco

smell in the house was unbearable'], [Subject + Verb + Object (NP): *el vecindario sentía el olor a tabaco de la casa*, 'the neighborhood noticed the tobacco smell of the house'] and [Subject + Verb + Prepositional Complement (Prep + NP): *Me enamoré del olor a campo de su ropa*, 'I fell in love with the country smell of their clothes']. This will allow us to generate sentence contexts with the most frequent valency patterns. Again, new sets of lexical prototypes will be created for the rest of slots of the sentence contexts on the basis of frequency queries in corpora and dictionaries.

As in the preceding working steps, it will be necessary to perform a manual assessment of the acceptability of a representative amount of output generated for each language. As a final result, users should be able to decide in a web interface the types of context they want to be displayed.

3 Conclusions

The combined methodology for data extraction based on the interoperability of different resources, as presented in this paper, will lead to the development of the MultiGenera prototype. Beyond this specific outcome, the project not only entails the description of lexical-semantic fields, but also the study of the noun as a valency carrier. To this end, a multi-layered analysis of syntagmatic and paradigmatic combination patterns and their distribution is being conducted for three different languages. Among potential applications of the results of MultiGenera, we may highlight, on the one hand, that the project will help to explore new ways for the semantic annotation of corpora using semantic categories and existing lexical knowledge ontologies such as wordnets. This might be interesting to improve several Natural Language Processing tasks, such as Natural Language Generation or rule-based Machine Translation. On the other hand, MultiGenera will also allow us to generate examples of valency patterns and argument structure realizations in three languages, which well may find application in language teaching and Lexicography.

References

Bond, F. & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. ACL 2013, Sofia, pp. 1352–1362.

Domínguez Vázquez, M.J. & Valcárcel Riveiro, C. (in press). PORTLEX as a multilingual and cross-lingual online dictionary. In M.J. Domínguez Vázquez, M. Mirazo Balsa, C. Valcárcel Riveiro (eds.) Studies on multilingual lexicography.

Domínguez Vázquez, M.J. (2011). Kontrastive Grammatik und Lexikographie: spanisch-deutsches Wörterbuch zur Valenz des Nomens. München: Iudicium.

Engel, U. (1995). Tiefenkasus in der Valenzgrammatik. In L. Eichinger, H.-W. Eroms (eds.) Dependenz und Valenz, Hamburg: Buske, pp. 53-65.

Gómez Guinovart, X. & Solla Portela, M.A. (2018). Building the Galician wordnet: methods and applications. In *Language Resources and Evaluation*, 52(1), pp. 317-339.

Gonzalez, A., Laparra, E. & Rigau, G. (2012). Multilingual Central Repository version 3.0. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul: ELRA

Gonzalez, A., Rigau, G. & Castillo, M. (2012). A Graph-Based Method to Improve WordNet Domains. In Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science. International Conference on Intelligent Text Processing and Computational Linguistics, Springer, Berlin, Heidelberg, pp. 17–28.

Mel'čuk, I. (2015). Semantics. From meaning to text, vol. 3, Amsterdam/Philadelphia: John Benjamins.

Mel'čuk, I. (2013). Semantics. From meaning to text, vol. 2, Amsterdam/Philadelphia: John Benjamins.

Niles, I. & Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03), CSREA Press, pp. 412–416.

Rodríguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertagna, F. & Roventini, A. (1998). The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. In Computers and the Humanities, 32, 117–152.

Sagot, B. & Fišer, D. (2008). Building a free French wordnet from multilingual resources. In Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08). Marrakech: ELRA.

Valcárcel Riveiro, C. & Domínguez Vázquez, M.J. (2016). Teste 'muerte': falantes a avaliar a aceitabilidade de frases nominais geradas artificialmente. Blog Post, Carlos Valcárcel Riveiro. Retrieved from https://carlosvalcarcel.net/2016/11/30/teste-muerte-falantes-a-avaliar-a-aceitabilidade-de-frases-nominais-geradas-artificialmente/ [28.03.2018]

Acknowledgements

The results of this work are related to the research project "Generación multilingüe de estructuras argumentales del sustantivo y automatización de extracción de datos sintáctico-semánticos", financed by the BBVA Foundation Grants for Scientific Research Teams 2017, and to the research project "Multilingual generator of noun argument structures with application in foreign language production", financed by the Spanish Ministry of Economy, Industry and Competitiveness (Scientific and Technical Excellence Research Program, FFI2017-82454-P).