Resource Interoperability: Exploiting Lexicographic Data to Automatically Generate Dictionary Examples

María José Domínguez Vázquez[1], Miguel Anxo Solla Portela[2], Carlos Valcárcel Riveiro[3]

[1] Department of English and German Philology and Galician Language Institute, University of Santiago de Compostela
[2] Department of English and German Philology, University of Santiago de Compostela
[3] Department of English, French and German Philology, University of Vigo
Email: majo.dominguez@usc.es, miguel.solla@usc.es, carlos.valcarcel.riveiro@uvigo.es

Abstract

This paper describes the different design and development stages of the MultiGenera and MultiComb prototypes for the multilingual automatic generation of dictionary examples that contain nominal argument patterns at the phrasal and sentence levels. The main objective of MultiGenera is the development of a simulator for the automatic generation of phrases in Spanish, German and French, which is based on the argument patterns of ten valency nouns. The second one, MultiComb, aims to automatically generate the phrasal and sentence contexts of the previously selected nouns in MultiGenera. In the present study we focus on the description of resource interoperability and a set of tools developed to support the methodology of both projects.

Keywords: Valency Dictionary; Argument Patterns; Natural Language Generation; WordNet; Semantics and Ontologies

1. Introduction

The advances in the automatic generation of the natural language have allowed the development of many applications following different methodologies, and thus it has been possible to generate many varied texts, from meteorological forecasts to song lyrics. However, in many cases the texts generated lack meaning or coherence. The *MultiGenera* and *MultiComb* projects were launched to help tackle these problems by exploring the potential of the information contained in valency dictionaries and take advantage of the opportunities offered by WordNet for lexical data extraction. This article presents the different steps taken in developing the tools and prototypes within these projects, focused on the automatic generation of noun phrases and their sentence contexts in Spanish, German and French.

The next section explains in more detail the core principles of the MultiGenera and MultiComb projects. Section 3 focuses on the main features of the PORTLEX dictionary and on how the workflow for this project led to the idea of developing

MultiGenera and MultiComb (for more information see Domínguez Vázquez, Lindemann & Valcárcel Riveiro, 2018). In section 4, the combined theoretical and methodological approaches for the automatic generation of linguistic data are explained. This section describes how prototypical lexical units are obtained for filling in argument slots. Furthermore, it presents the process of lexical expansion, a phase prior to automatic generation, and the role of WordNet ontologies for this purpose. The functionalities and uses of the developed tools (APIs, LEMATIZA, COMBINA and XERA) are also presented in this section. Finally, a brief summary of the main ideas discussed will serve as the conclusion of this work.

2. General framework

The main goal of the MultiGenera project is to develop a tool for automatically generation of nominal phrases in Spanish, German and French. Some pre-project tests (Valcárcel Riveiro & Domínguez Vázquez, 2016) led us to the idea that the semantic acceptability of automatically generated noun phrases may be improved by providing enriched phrasal and sentence contexts. This assumption is actually at the basis of the MultiComb project, which aims to offer a simulator for creating acceptable sentence contexts for noun phrases in the three languages involved: Spanish, German and French. It is therefore a question of progressing from a valency noun with its different arguments to a sentence that contains it.

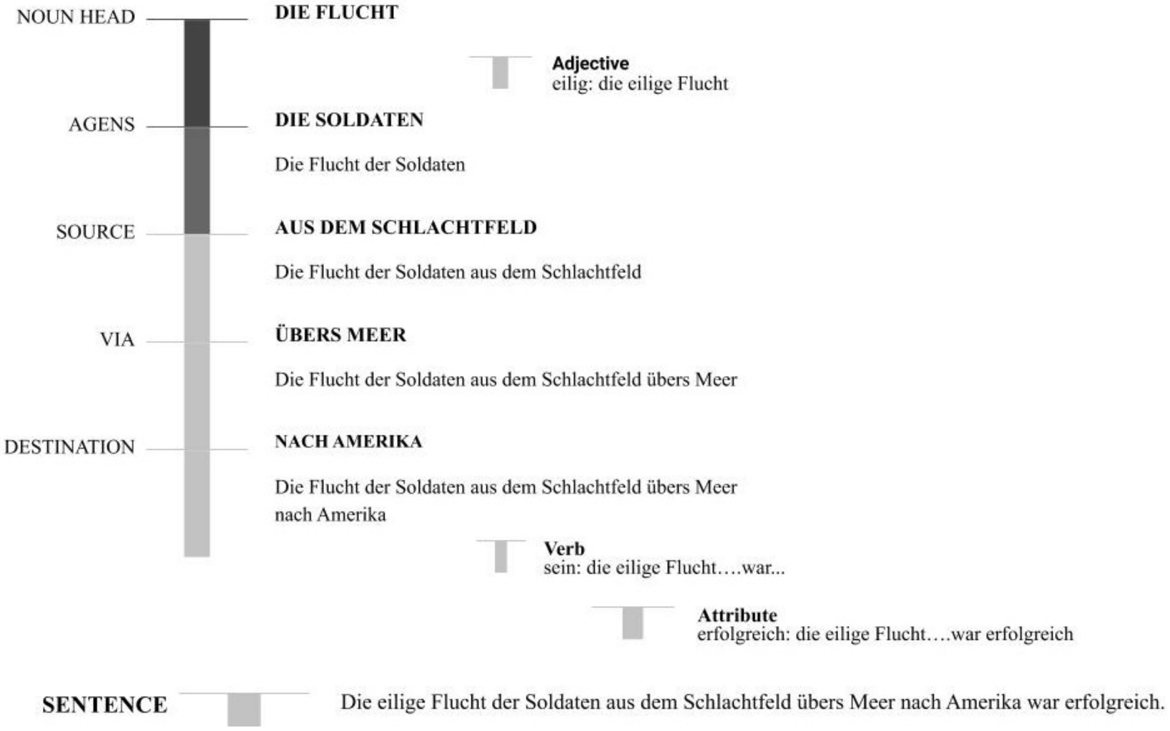

NOUN HEAD
DIE FLUCHT
Adjective
eilig: die eilige Flucht
AGENS
DIE SOLDATEN
Die Flucht der Soldaten
SOURCE
AUS DEM SCHLACHTFELD
Die Flucht der Soldaten aus dem Schlachtfeld
VIA
ÜBERS MEER
Die Flucht der Soldaten aus dem Schlachtfeld übers Meer
DESTINATION
NACH AMERIKA
Die Flucht der Soldaten aus dem Schlachtfeld übers Meer nach Amerika
Verb
sein: die eilige Flucht….war...
Attribute
erfolgreich: die eilige Flucht….war erfolgreich
SENTENCE
Die eilige Flucht der Soldaten aus dem Schlachtfeld übers Meer nach Amerika war erfolgreich.

Figure 1: Progression in building examples[1]

[1] Literal translation of the example in Table 1: ‘The hasty escape of the soldiers from the battlefield by sea to the Amerika was successful’.

The development of both projects is fed by different theoretical and methodological approaches from different linguistic theories, such as Valency Grammar, Prototypes Theory, Meaning-to-Text Theory and Natural Language Processing (NLP). Furthermore, our combined method utilizes i) the automatic extraction of data from NLP resources, ii) the analysis of corpora, co-occurrence databases and wordnets, iii) as well as the outcoming evaluation produced by both generators.

This paper presents a way of exploiting existing lexicographic information (see section 3) to generate new lexicographic data based on custom-made tools (MultiTools[2]) and on resource interoperability. Specifically, the following tools have been developed in the current phase:

1) Three query APIs, one for each language[3], were designed with the aim of extracting lexical data from queries pointing to the semantic relations of WordNet and to the ontologies linked to the synsets in the EuroWordNet model (see 4.3.1). They provide the results in a standard data exchange format (JSON).
2) LEMATIZA[4] analyses exported documents from corpora and provides the lemma of the inflected form of each argument. Each lemma is linked to all the possible queries to the API for the corresponding language. This tool significantly reduces time spent in formulating queries with a semi-automatic query selection (see 4.3.2).
3) Another application, COMBINA[5] makes it possible to combine or crosscheck the results of several API queries. Most of the time, the typology of classes available with simple queries does not conform to an 'ontology' of classes based on linguistic semantics. However, a combination of queries offers an enormous variety of possibilities and manages to fine-tune the results with great precision. In addition, these new classes are easily reusable (and even perhaps implementable as a new ontology linked to wordnets) (see 4.3.3).
4) A prototype of a generator of noun phrases, XERA[6], is also being developed for the three languages (see 4.4).

In relation to the foregoing it should be noted that exploring data bootstrapping from NLP resources is interesting for MultiGenera and MultiComb, and therefore for the resources on which they are based. Resource interoperability is understood here in two directions:

[2] http://portlex.usc.gal/develop/

[3] The API functionalities are described in the following links, from which queries can also be launched. Spanish API: http://portlex.usc.gal/develop/es/api/; French API: http://portlex.usc.gal/develop/fr/api/; German API: http://portlex.usc.gal/develop/de/api/.

[4] http://portlex.usc.gal/develop/lematiza/

[5] http://portlex.usc.gal/develop/combina.php

[6] http://portlex.usc.gal/develop/xera.php

1) The use of data from, for example, WordNet ontological features, PORTLEX's argument patterns (see Section 3) and the dictionaries from the FreeLing tagger (Padró, 2011) for the development of our generators. so that the inflector, although it is also custom-made, reuses FreeLing's dictionaries.

2) The use of our generators and tools to improve other resources or design new ones. Thus, for example, resources on lexical selections are offered in JSON format so that they can be used directly by other applications. A further illustration of the intended interoperability is the possible exploitation of our APIs and tools, such as COMBINA and LEMATIZA.

3. The PORTLEX dictionary as a starting point for developing MultiGenera and MultiComb

PORTLEX[7] is an online valency dictionary of noun phrases with application in language production. It compiles multilingual data in German, Galician, Spanish, Italian and French. The main features of this resource are:

(1) **valency based** (Engel, 2009): PORTLEX provides detailed information on the nominal phrase from the point of view of valency grammar. This dictionary primarily concerns deverbal (EVALUACIÓN 'evaluation', INVESTIGACIÓN 'research', etc.) and deadjectival nouns (SINCERIDAD 'sincerity', TRANQUILIDAD 'tranquillity', etc.), but also non-derivative nouns that present valency patterns such as PROBLEMA 'problem', GANA 'desire, craving', among others. The specific arguments and semantic roles constitute first-order elements in the entries microstructure. On the one hand, a series of roles are defined to identify the semantic function of the nouns' arguments (e.g. 'that which performs an action', 'that which is affected', etc.) as well as their syntactic function (*subiectivus, obiectivus*, etc.). On the other hand, the semantic description also resorts to a list of semantic features ('animate', 'institution', 'object', 'situation', etc.) associated with the valency arguments and present in the different formal realizations of each argument.

(2) **online** (Klosa, 2013; Müller-Spitzer, 2014) and **semi-collaborative** (Abel & Meyer, 2013; Melchior, 2014): Regarding its medial features, this dictionary was developed as an online and continuously updated resource based on hypertextualization, user interaction and combined access. It is not a finished work, but is constantly updated thanks to its semi-collaborative nature.

(3) **modular**, **multilingual** and **cross-lingual** (Domínguez Vázquez & Valcárcel Riveiro, 2019; Gouws, 2014): Domínguez Vázquez & Valcárcel Riveiro (2019: 140)

[7] http://portlex.usc.gal/portlex/

> describe these features as follows: “The PORTLEX dictionary covers five languages contrasted with each other. Indeed, its database is designed to include more languages. It contains a specific module for each language in which data relating to each one of them is stored. These modules are linked to each other through a mother dictionary (Gouws, 2014) where Spanish is the pivot language. This allows the alignment of the data of each language and enables their contrastive display according to the user’s needs. In this way, PORTLEX can be defined not only as a multilingual dictionary, but above all also as a cross-lingual dictionary [...]”.

A valency dictionary should provide syntactic and semantic information that helps its users to improve their linguistic production in a foreign language. Therefore, any valency dictionary must describe the different argument realizations of a lexeme, their combining rules and the syntactic-semantic restrictions attached to them, since its aim is to provide users with a complete and detailed description of argument patterns (Domínguez Vázquez, 2018). In order to get a broad dataset PORTLEX relied on corpora for the different languages described in the dictionary and thoroughly analysed them. The examination of the compiled corpus-data allowed the observation that many extracted examples or surface realizations did not meet the requirements of a valency dictionary and, in this sense, we encountered difficulties related to the following issues:

i. The time-consuming corpus-based compilation of all the noun surface realizations. In this case, the search for certain realizations functioning as noun complements, such as adjectives and compounds, is very time consuming, since they are either scarcely represented in the large corpora used or are not found in them even though they do exist.

ii. The tedious description of the noun argument patterns, i.e. the compilation of all possible combinations and syntactic-semantic restrictions for each argument along with their different surface realizations in the five languages of the PORTLEX dictionary. The combination patterns of the German noun FLUCHT ‘flight’/‘escape’ well exemplifies such cases, since it presents four arguments: A1: argument with the role ‘that which performs the action’, A2: Argument with the role ‘origin’, A3: Argument with the role ‘transit’ and A4: Argument with the role ‘destination’.

A1	A2	A3	A4
1. Genitive	1. von + dative	1. durch + accusative	1. in + accusative
2. von + dative	2. von ... aus	2. über + accusative	2. auf + accusative
3. Adjective	3. aus + dative	3. via	3. nach + dative
4. Compound	4. Compound		4. zu + dative
			5. bis + preposition + dative

Figure 2: Arguments and surface realizations of the German noun FLUCHT.

In its current state the dictionary describes 61 patterns for the noun FLUCHT, such as the following:

16 monoargumental patterns

$A1_1$ = Die Flucht der Tiere
$A1_2$ = Die Flucht von 231 Migranten
$A1_3$ = Die väterliche Flucht
$A1_4$ = Die Einwohnerflucht
$A2_3$ = Die Flucht aus Spanien
$A2_4$ = Die Stadtflucht

31 biargumental patterns

$A1_1 + A2_1$ = Die Flucht der Familie aus Spanien
$A1_4 + A2_1$ = Die Tierflucht aus dem Zoo
$A1_1 + A3_1$ = Die Flucht der Gefangenen durch den Wald
$A1_2 + A4_3$ = Die Flucht von DDR-Bürgern nach West-Berlin
$A2_3 + A3_2$ = Die Flucht aus Prag über Salzburg
$A3_1 + A2_3$ = Die Flucht durch einen Tunnel aus dem Gerichtssaal
$A3_3 + A4_3$ = Die Flucht via Jugoslawien nach Österreich
$A4_3 + A1_2$ = Die Flucht nach Amerika von Carl Schurz

13 triargumental patterns

$A1_4 + A2_1 + A4_4$ = Die Lehrerflucht von öffentlichen zu privaten Schulen
$A2_3 + A3_1 + A4_1$ = Die Flucht aus der Erdgeschosswohnung durch das Fenster in den Innenhof.
$A1_2 + A3_2 + A4_3$ = Die Flucht von EU-Bürgern über Thailand nach Japan

1 Tetrargumental pattern

$A1_2 + A2_3 + A3_1 + A4_5$ = Die Flucht von Räubern aus China durch Europa bis in die Schweiz

Figure 3: Argument patterns of the German noun FLUCHT.

As Figure 3 shows, the main difficulties arise in describing the combinatorial arguments, i.e. the interaction of each involved argument in all their realizations and distribution possibilities.

iii. Corpus-extracted data often do not suit the requirements of a valency dictionary. This is mainly due to the fact that most corpora are not semantically tagged. This is a real concern, as the head of an argument, which represents a certain semantic role (Engel, 1996), must present specific semantic features accordingly, regardless of its formal realization (prepositional phrase, adjective phrase, apposition, compound name, etc.). As shown in Figure 2, for example, the German lexeme FLUCHT has four different surface realizations for its agent complement (A1). The use of a compound noun 'agent'+FLUCHT (*Die Einwohnerflucht*) is one of these possible realizations. However, a query on the German web 2013 (deTenTen 13) using Sketch Engine[8] retrieves all kinds of compounds (*Die Weiterflucht* or *die Berufsflucht*), since these can't be semantically filtered. In fact, most extracted compound nouns do not contain any agent in their first element. A syntactic-semantic analysis of the 100 most frequent lemmas in the mentioned search (Figure 4) shows that a semantic analysis leads us to reject many of them, and this is because the agent of FLUCHT has to feature the semantics characteristics 'human', 'animal' or 'vehicle'.

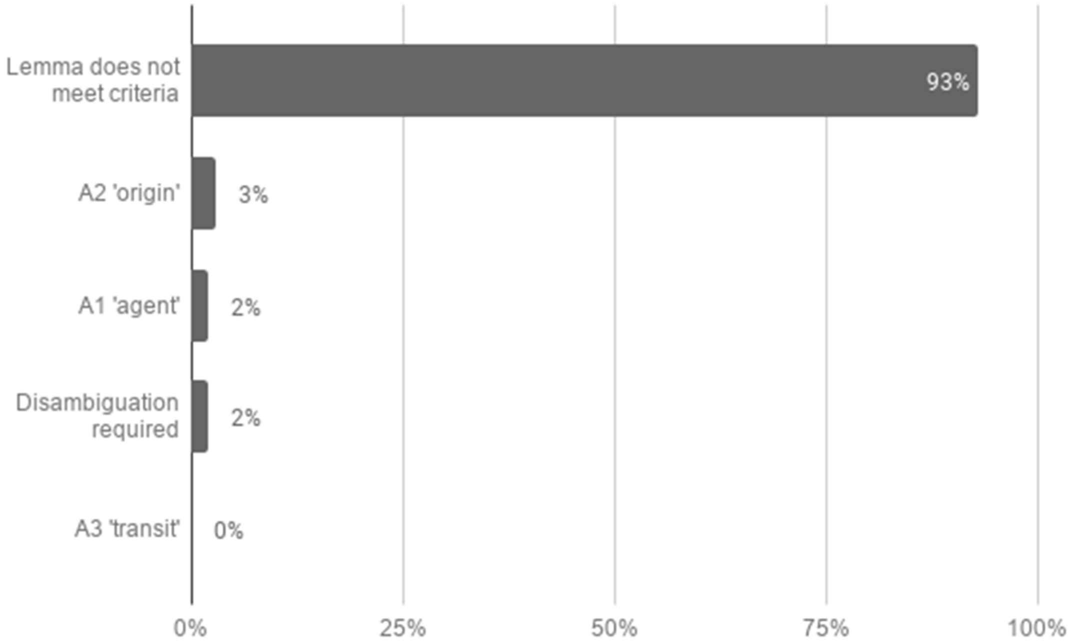


Figure 4: Semantic analysis of the compound nouns retrieved for FLUCHT (deTenTen13)

These cases, in which two or more noun arguments present the same formal realization, are quite frequent. Since we obtain argumental realizations from corpora thanks to their grammatical annotation, in many cases the results show occurrences that are formally similar to the argumental realization that we are searching for, but that actually correspond to another, different semantic role. Thus, very often observing the semantic features of a corpus realization is the only way to determine to which semantic argument it belongs. This means that a human review of the entire list of a query results is necessary to find the examples which can represent a specific semantic role.

[8] https://www.sketchengine.eu/

And it is precisely here where MultiGenera's strength lies, because this project tackles not only the semantic roles of arguments, but also the distinctive semantic features shared within the lexical paradigms involved in their slot-filling. For this reason, it is not enough to pick up the lexical units retrieved by queries in large corpora (it is not even always representative due to metaphorical uses of the nominal head or their arguments, context dependence for interpretation). The project aims to solve this problem by first identifying the semantic prototypes involved in the roles of the arguments. Ultimately, the purpose is thus the creation of semantically coherent paradigms for the generation of natural language that are independent of context[9].

4. MultiGenera and MultiComb: theoretical and methodological approaches

4.1 Starting Point

We start from a combined approach for the collection and analysis of data on noun phrases for Spanish, German and French (see section 1). This procedure allows combining valency grammar, the lexical prototype theory, semantic classes and natural language processing (information retrieval and extraction, as well as natural language generation). The automatic generation of the nominal phrase and its arguments relies specifically on a combined method, which is based on the following methodological phases shown in Figure 5:

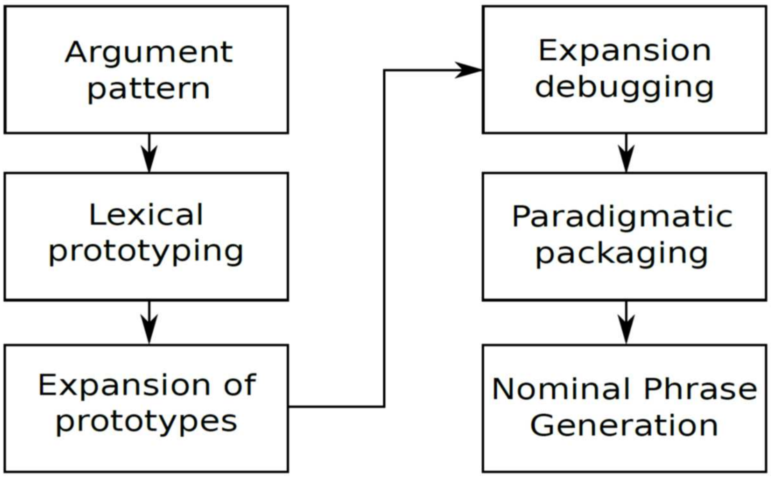


Figure 5: Combined method phases

[9] MultiComb project deals with the context generation.

In the following sections we will focus on the argument pattern and the lexical prototyping phases (4.2), as well as on the procedure for the prototypes expansion (4.3) and the generation of nominal phrases (4.4).

4.2 Argument pattern and lexical prototyping

The PORTLEX dictionary is used to obtain syntactic and semantic patterns of noun arguments in Spanish, German and French:

Argument s	**A1**	**A2**	**A3**	**A4**
Semantic role	'that which performs an action'	Location: origin	Location: transit	Location: destination
Semantic feature	[animate]	[place], [locality], [territory]	place], [locality], [territory]	[place], [locality], [territory]

Table 1: Argument structure of A1 und A2 and semantic features for the German noun FLUCHT.

Argument patterns in PORTLEX provide the parameters for the route queries in Sketch Engine's corpora. There queries are designed to identify lexical units that could fill the argument slots of the nouns selected. To illustrate it we will provide the following example with FLUCHT: we search precisely for the slot-filling nouns for A2 (semantic role 'origin'; see Table 1) in coappearance with the preposition *aus* (Table 2). A detailed semantic examination of the examples obtained from CQL[10] queries is carried out following a frequency criterion. Lexical units such as *DDR* 'GDR', *Ghetto* 'ghetto', *Troja* 'Troja', *Haus* 'home', *Frankreich* 'France', *Ost-Berlin* 'East Berlin', *Ostgebieten* 'eastern territories' and *Kriegsgefangenenlager* 'POW camp' appear frequently in the Sketch Engine corpus as examples for A2-Nouns and thus are, according to our methodological approach, prototypical slot-candidates. The identification of these lexical prototypes makes it possible to define the main semantic classes involved in the slot-filling of each noun argument. This proceeding enables, from these lexical prototypes, to propose the main semantic classes from among the categories of a custom-made linguistic ontology with semantic classes (Table 2):

[10] Corpus Query Language (see https://www.sketchengine.eu/documentation/corpus-querying/)

FLUCHT
aus+dative +

Lexical prototypes	1st Level	2nd Level	3rd Level	4th Level
Warschauer Ghetto	situation	location	territory	
Haus	situation	location	building	
Kriegsgefangenenlager	situation	location	building	
Wohnung	situation	location	building	
Troja	situation	location	locality	proper name
Ost-Berlin	situation	location	locality	proper name
Venedig	situation	location	locality	proper name
Frankreich	situation	location	territory	proper name
Deutschland	situation	location	territory	proper name
Italien	situation	location	territory	proper name

Table 2: Example of semantic annotation of lexical prototypes for the argument pattern A23 FLUCHT + aus.

By prototyping we get to establish not only the most representative semantic classes of the different argument patterns, but also the constraints involved in the lexical selection of the focal pattern, such as in the following example for the semantic role ‘source’ of the argument pattern $A2_3$ (FLUCHT aus + dative):

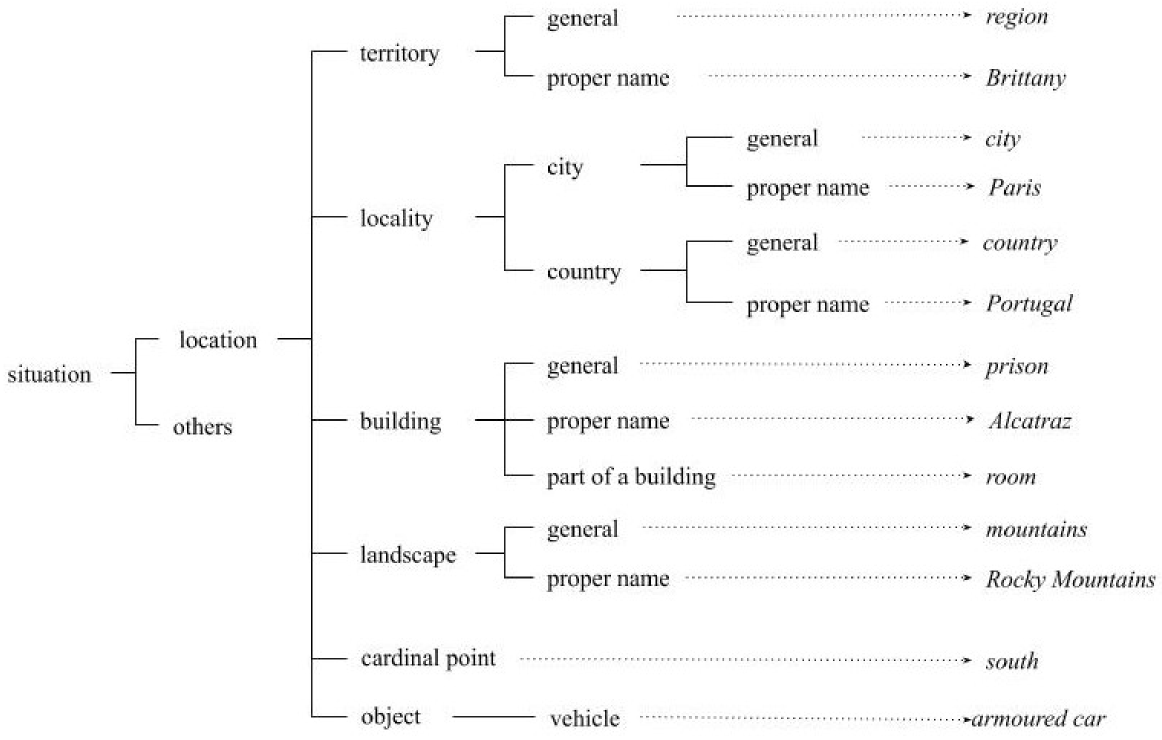


Figure 6: Prototypical semantic classes of FLUCHT + aus

4.3 Expansion of prototypes

4.3.1 Resorting to WordNet

The semi-automatic extraction of lexical candidates for the paradigmatic axis of each argument relies on the fact that the synsets of the wordnets following the EuroWordNet model of the Multilingual Central Repository (MCR)[11] (González Agirre & Rigau, 2013) are associated with semantic or cognitive features categorized in different ontologies. In particular, we are dealing with Suggested Upper Merged Ontology[12] (SUMO) (Niles & Pease, 2001), Top Concept Ontology[13] (Top) (Álvez et al., 2008), WordNet Domains[14] (Bentivogli et al., 2004), Basic Level Concept (Izquierdo et al., 2007) and Epinonyms (Gómez Guinovart & Solla Portela, 2018). Therefore, it is necessary to identify the categories that resort to a concrete wordnet and enable us to fill in the valency slots according to the required semantic feature. For this, besides the ontologies already mentioned, we also use the semantic primes (Miller et al., 1990), i.e., the semantic primitives that organize the lexicographic files of nouns in WordNet, and even the semantic relations among synsets.

Nevertheless, the difficulty in establishing these connections arises from the fact that the cognitive organization of the ontological classifications in the wordnets of the MCR and Galnet[15] (Galician WordNet development interface) do not exactly follow a fully adequate organization for the linguistic description required for MultiGenera. In spite of this, many of the semantic classes defined for our project also constitute categories or general classes in ontologies that are already present in the MCR, such as Top, SUMO, WordNet Domains or Epinonyms. The difficulty consists, therefore, in establishing the appropriate channels for obtaining lexical repertoires with finer semantic granularity to fill in the argument slots of each surface realization. But, in addition, the decision to resort to WordNet has entailed a series of initial tasks, since at the beginning of MultiGenera and MultiComb only Spanish had a wordnet linked to the aforementioned ontologies, as part of the MCR. Thus, the first step undertaken was the creation of databases for French and German. This was done by extracting the alignment between lexical variants and identifying offsets of the meaning from the WordNet Libre du Français[16] (WOLF) (Sagot & Fišer, 2008) and with data from the Extended Open

[11] http://adimen.si.ehu.es/web/MCR

[12] http://www.adampease.org/OP/

[13] http://globalwordnet.org/gwa/ewn_to_bc/ewnTopOntology.htm

[14] http://wndomains.fbk.eu/

[15] http://sli.uvigo.gal/galnet/index.php?lg=en. We link to the multilingual web interface of the Galician wordnet to explore the synsets.

[16] https://gforge.inria.fr/projects/wolf/

Multilingual WordNet[17] (Bond & Foster, 2013). Both have been made available on the Galnet interface after being converted to the EuroWordNet format of the MCR. In this way, the links with the categories of the ontologies discussed above are available to operate in the three languages of the project. Since syntactic arguments perform semantic roles with their respective ontological-semantic features, we can turn to a lexicon, in this case wordnets, to fill in the argument slots of the selected nouns with lexical units. Expansions of the lexical prototypes described earlier can be made by connecting their semantic classes with the categories of ontologies linked to WordNet in combination with other selection criteria based on the internal structure of this lexical-semantic network. In such a way, through queries in the wordnets, we obtain series of synsets with a meaning that meets the semantic requirements of the lexical paradigms of a noun argument. From these synsets we extract the variants of each language to integrate them into the lexical paradigm of the argument concerned. These connections between semantic classes and WordNet ontological categories can be made using two custom-made designed tools: LEMATIZA and COMBINA.

Figure 7 illustrates that the Semantic Prototype Class (SPC) [situation, location, locality, proper name] is connected with three categories from three different ontologies linked to the wordnets by intersecting the synsets that share these categories.

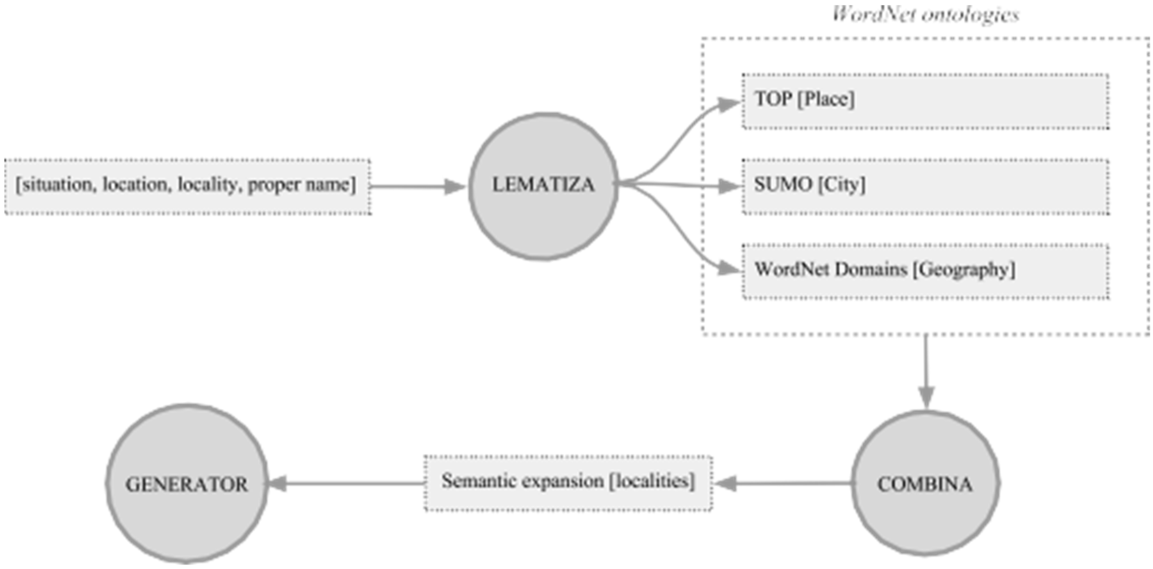

WordNet ontologies
TOP [Place]
SUMO [City]
WordNet Domains [Geography]
[situation, location, locality, proper name]
LEMATIZA
GENERATOR
Semantic expansion [localities]
COMBINA

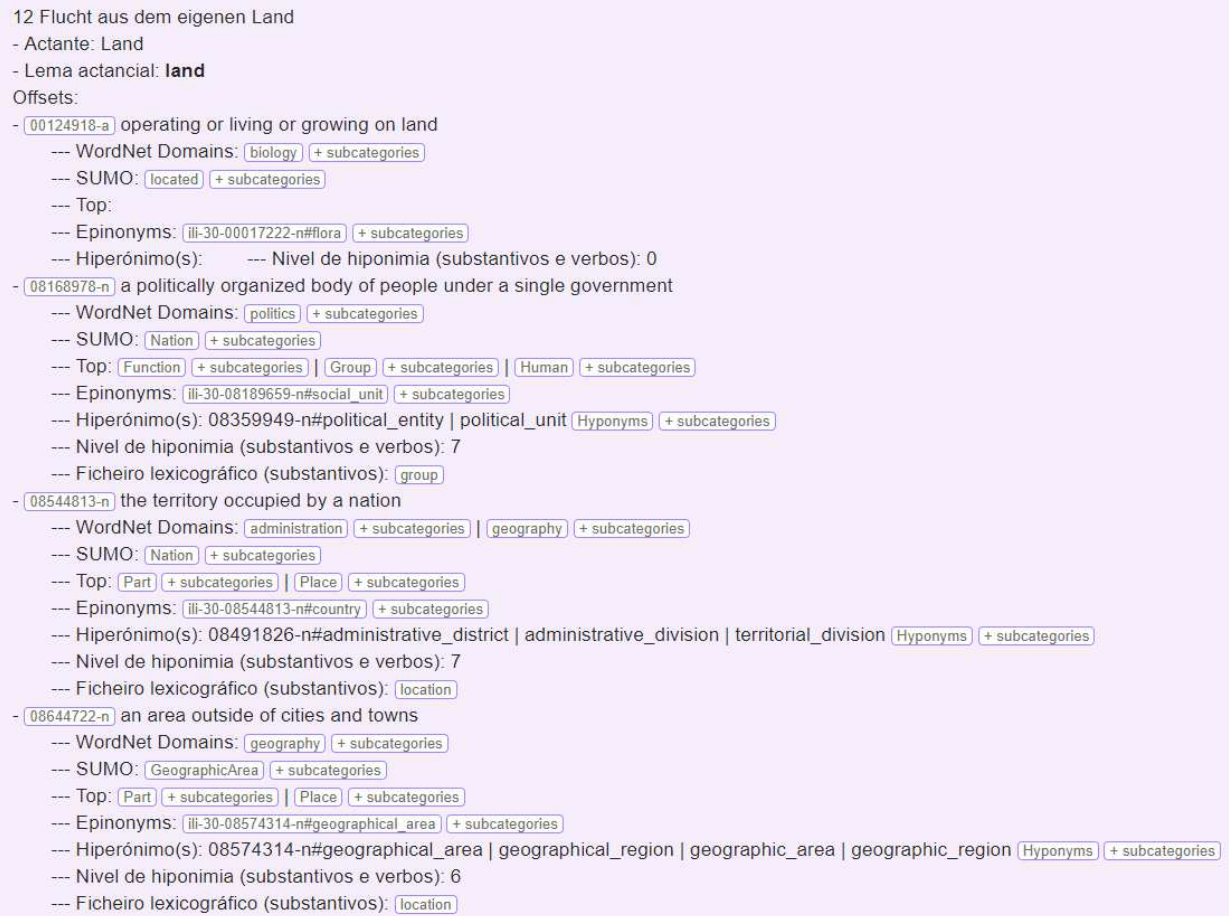
Figure 7: Tools for semantic analyses and expansion by using the wordnets

This procedure allows us to obtain a lexical selection or paradigm with the same semantic characteristics of the initial lexical-semantic prototype. The debugging of the lexical expansion establishes the paradigmatic axis that supports the lexical selection in the automated generation of phrasal contexts. Below the functionalities of the LEMATIZA (4.3.2), COMBINA (4.3.3) and XERA (4.4) will be explained in more detail.

[17] http://compling.hss.ntu.edu.sg/omw/summx.html

4.3.2 LEMATIZA

LEMATIZA aims to ease more appropriate queries in the APIs (see section 2). This robust tool allows introducing both concordances and frequency lists, as exported from Sketch Engine, in any of the three languages involved. LEMATIZA returns lemmas from the inflected forms of argument realizations retrieved from CQL queries in Sketch Engine. Each resulting lemma is searched, in turn, in the WordNet of the corresponding language and the output shows each of the synsets in which it is present. In addition, and importantly, this tool provides links to API queries pointing to the ontological categories of each synset, as well as to internal queries to its direct hypernym and hyponyms (see Figure 8) and all its hyponymic descendants. Since LEMATIZA offers links for all the synsets of a lemma, a process of manual disambiguation needs to be carried out to identify the meaning according to that specific usage in the corpus. Disambiguated query links are combined to get the lexical selection for each argument. Moreover, this also allows us to validate the semantic categories of the ontology that we build in order to semantically organize, structure and, when possible, reuse all the lexical selections of our projects.

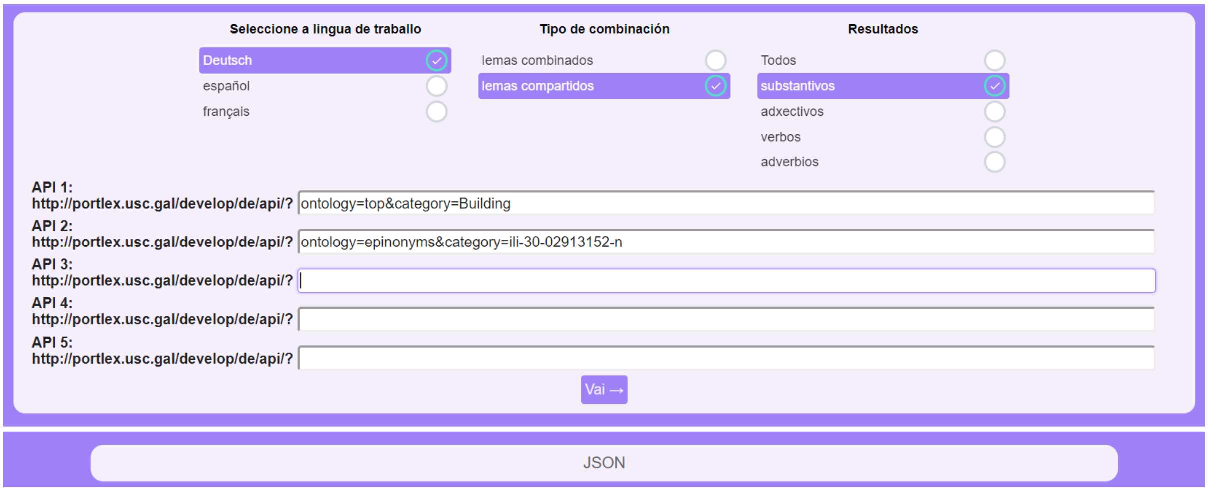


```
12 Flucht aus dem eigenen Land
- Actante: Land
- Lema actancial: land
Offsets:
- 00124918-a operating or living or growing on land
      --- WordNet Domains: biology + subcategories
      --- SUMO: located + subcategories
      --- Top:
      --- Epinonyms: ili-30-00017222-n#flora + subcategories
      --- Hiperónimo(s):      --- Nivel de hiponimia (substantivos e verbos): 0
- 08168978-n a politically organized body of people under a single government
      --- WordNet Domains: politics + subcategories
      --- SUMO: Nation + subcategories
      --- Top: Function + subcategories | Group + subcategories | Human + subcategories
      --- Epinonyms: ili-30-08189659-n#social_unit + subcategories
      --- Hiperónimo(s): 08359949-n#political_entity | political_unit Hyponyms + subcategories
      --- Nivel de hiponimia (substantivos e verbos): 7
      --- Ficheiro lexicográfico (substantivos): group
- 08544813-n the territory occupied by a nation
      --- WordNet Domains: administration + subcategories | geography + subcategories
      --- SUMO: Nation + subcategories
      --- Top: Part + subcategories | Place + subcategories
      --- Epinonyms: ili-30-08544813-n#country + subcategories
      --- Hiperónimo(s): 08491826-n#administrative_district | administrative_division | territorial_division Hyponyms + subcategories
      --- Nivel de hiponimia (substantivos e verbos): 7
      --- Ficheiro lexicográfico (substantivos): location
- 08644722-n an area outside of cities and towns
      --- WordNet Domains: geography + subcategories
      --- SUMO: GeographicArea + subcategories
      --- Top: Part + subcategories | Place + subcategories
      --- Epinonyms: ili-30-08574314-n#geographical_area + subcategories
      --- Hiperónimo(s): 08574314-n#geographical_area | geographical_region | geographic_area | geographic_region Hyponyms + subcategories
      --- Nivel de hiponimia (substantivos e verbos): 6
      --- Ficheiro lexicográfico (substantivos): location
```

Figure 8: Screenshot (incomplete) of the data retrieved from LEMATIZA

4.3.3 COMBINA

For its part, the COMBINA tool has been developed with the purpose of integrating the API results more accurately. It combines the data from different API queries in the same language, either to add the data from one query to those of another or others (through the combined lemmas option) or to obtain the intersection of the results from different queries (shared lemmas). Figure 9 shows a COMBINA search for German lexemes that belong to the class 'Buildings'. An example of the results is shown in Table 3.



Seleccione a lingua de traballo	Tipo de combinación	Resultados
☑ Deutsch	○ lemas combinados	○ Todos
○ español	☑ lemas compartidos	☑ substantivos
○ français		○ adxectivos
		○ verbos
		○ adverbios

API 1:
http://portlex.usc.gal/develop/de/api/? ontology=top&category=Building

API 2:
http://portlex.usc.gal/develop/de/api/? ontology=epinonyms&category=ili-30-02913152-n

API 3:
http://portlex.usc.gal/develop/de/api/?

API 4:
http://portlex.usc.gal/develop/de/api/?

API 5:
http://portlex.usc.gal/develop/de/api/?

Figure 9: Screenshot of COMBINA

74 02977936-n Kasino	81 03007130-n Kirche	88 03078506-n Kommunikationszentrum
75 02984203-n Kathedrale	82 02820798-n Klapsmühle	89 03089753-n Konferenzzentrum
76 02984061-n Kathedrale	83 03043274-n Klinik	90 03092314-n Konservatorium
77 03032252-n Kino	84 02667576-n Kloster	91 03093427-n Konsulat
78 03028079-n Kirche	85 03054311-n Klubhaus	92 03093427-n Konsulatgebäude
79 02984061-n Kirche	86 04018399-n Kneipe	93 03540595-n Krankenhaus
80 02984203-n Kirche	87 03056288-n Kohlenkeller	94 03043274-n Krankenhaus

Table 3: Results retrieved from COMBINA by crossing API queries.

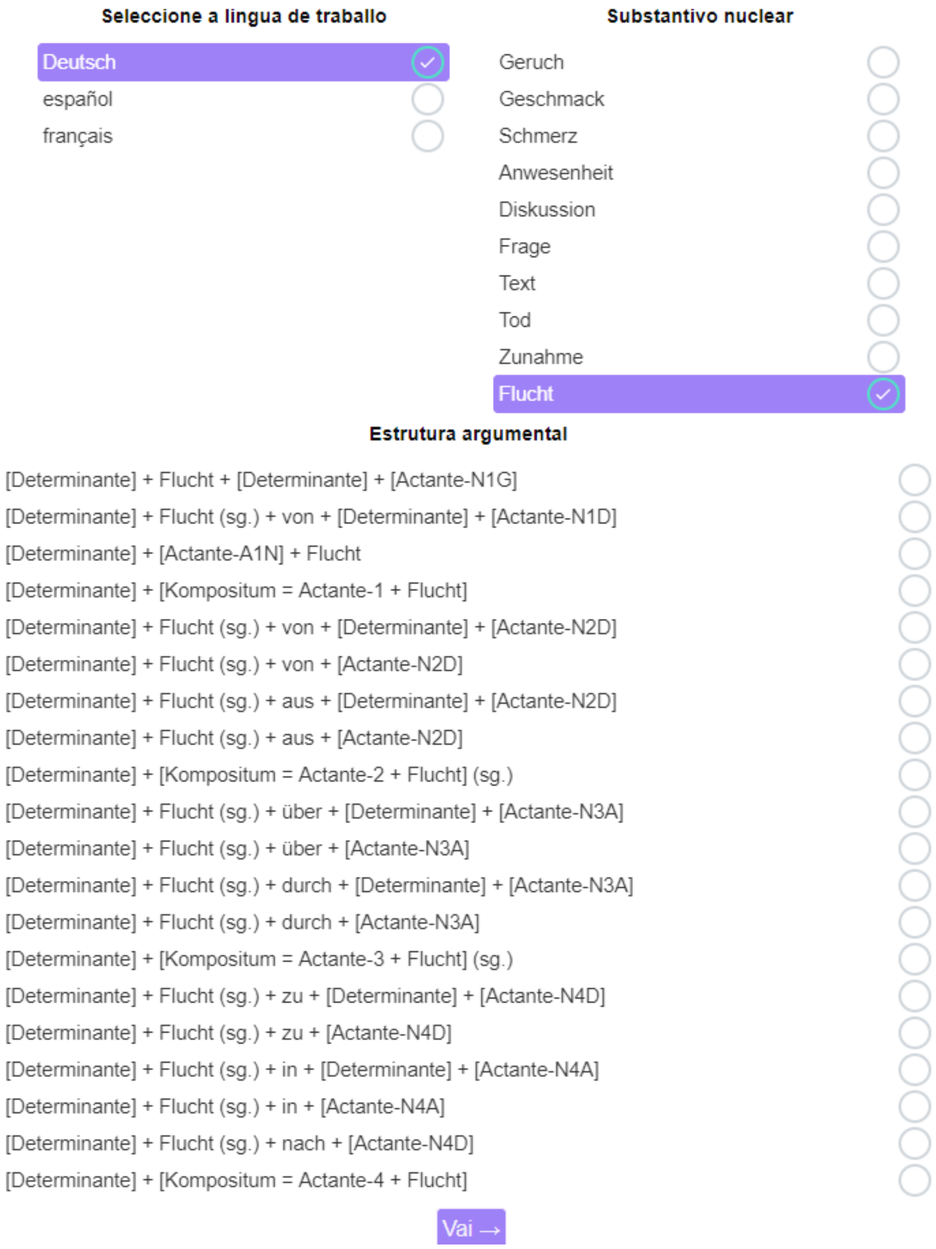
The results are provided in text format, but also in JSON so that they can be used directly by other applications (such as the prototype generator of MultiGenera, XERA). The debugging of the results constitutes the expanded lexical paradigms used for the automated generation of noun phrases.

4.4 Generation of the nominal phrase: phrasal and sentence context

All these previous steps lead to the design of the generator prototype for noun patterns, XERA[18] (see Figure 10). This tool generates nominal phrases using packaged lexical files built from the results of COMBINA searches. In query mode, it currently uses direct queries to an API or results from COMBINA in JSON format as input for lexical selections. The entire process is performed in real-time. Specific inflectors have been developed for each language, which provide the appropriate form for each context; that is, the inflection of case (only in German), gender and number for determinants, nouns (and the compounds argument + nucleus in the case of German) and adjectives (in German with formal variation depending on the determination, case and gender of the noun they accompany). The code that produces the inflected forms reuses the dictionaries[19] of the well-known tagger FreeLing. The presence of each lemma is verified and inflected forms are obtained by checking the morphosyntactic tags from the corresponding dictionary. In addition, in the case of German, at the moment we also run FreeLing so that it can, sometimes, offer the division into primary lemmas when compound forms are provided from a German wordnet. When the elements are inflected, the concordances and possible restrictions on the usage of all the words in the phrase are verified. The specific contractions of each language are carried out by means of functions that were specifically developed for this purpose.

[18] A more user-friendly interface will be designed in a later phase.

[19] See https://github.com/TALP-UPC/FreeLing/blob/master/COPYING



Seleccione a lingua de traballo

- Deutsch ✓
- español ○
- français ○

Substantivo nuclear

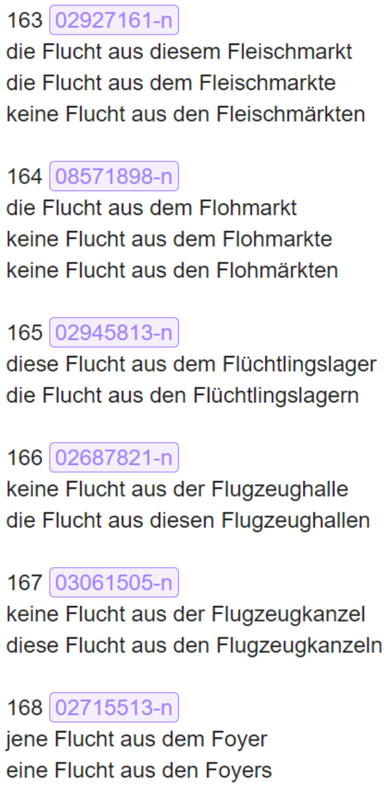
- Geruch ○
- Geschmack ○
- Schmerz ○
- Anwesenheit ○
- Diskussion ○
- Frage ○
- Text ○
- Tod ○
- Zunahme ○
- Flucht ✓

Estrutura argumental

- [Determinante] + Flucht + [Determinante] + [Actante-N1G] ○
- [Determinante] + Flucht (sg.) + von + [Determinante] + [Actante-N1D] ○
- [Determinante] + [Actante-A1N] + Flucht ○
- [Determinante] + [Kompositum = Actante-1 + Flucht] ○
- [Determinante] + Flucht (sg.) + von + [Determinante] + [Actante-N2D] ○
- [Determinante] + Flucht (sg.) + von + [Actante-N2D] ○
- [Determinante] + Flucht (sg.) + aus + [Determinante] + [Actante-N2D] ○
- [Determinante] + Flucht (sg.) + aus + [Actante-N2D] ○
- [Determinante] + [Kompositum = Actante-2 + Flucht] (sg.) ○
- [Determinante] + Flucht (sg.) + über + [Determinante] + [Actante-N3A] ○
- [Determinante] + Flucht (sg.) + über + [Actante-N3A] ○
- [Determinante] + Flucht (sg.) + durch + [Determinante] + [Actante-N3A] ○
- [Determinante] + Flucht (sg.) + durch + [Actante-N3A] ○
- [Determinante] + [Kompositum = Actante-3 + Flucht] (sg.) ○
- [Determinante] + Flucht (sg.) + zu + [Determinante] + [Actante-N4D] ○
- [Determinante] + Flucht (sg.) + zu + [Actante-N4D] ○
- [Determinante] + Flucht (sg.) + in + [Determinante] + [Actante-N4A] ○
- [Determinante] + Flucht (sg.) + in + [Actante-N4A] ○
- [Determinante] + Flucht (sg.) + nach + [Actante-N4D] ○
- [Determinante] + [Kompositum = Actante-4 + Flucht] ○

Figure 10: Example of argument patterns on the generator interface

The following screenshot shows the automatic generation for a search of the type “buildings you can flee from”, expressed in German with the preposition *aus*.



163 02927161-n
die Flucht aus diesem Fleischmarkt
die Flucht aus dem Fleischmarkte
keine Flucht aus den Fleischmärkten

164 08571898-n
die Flucht aus dem Flohmarkt
keine Flucht aus dem Flohmarkte
keine Flucht aus den Flohmärkten

165 02945813-n
diese Flucht aus dem Flüchtlingslager
die Flucht aus den Flüchtlingslagern

166 02687821-n
keine Flucht aus der Flugzeughalle
die Flucht aus diesen Flugzeughallen

167 03061505-n
keine Flucht aus der Flugzeugkanzel
diese Flucht aus den Flugzeugkanzeln

168 02715513-n
jene Flucht aus dem Foyer
eine Flucht aus den Foyers

Figure 11: Screenshot of XERA: automatically generated noun phrases

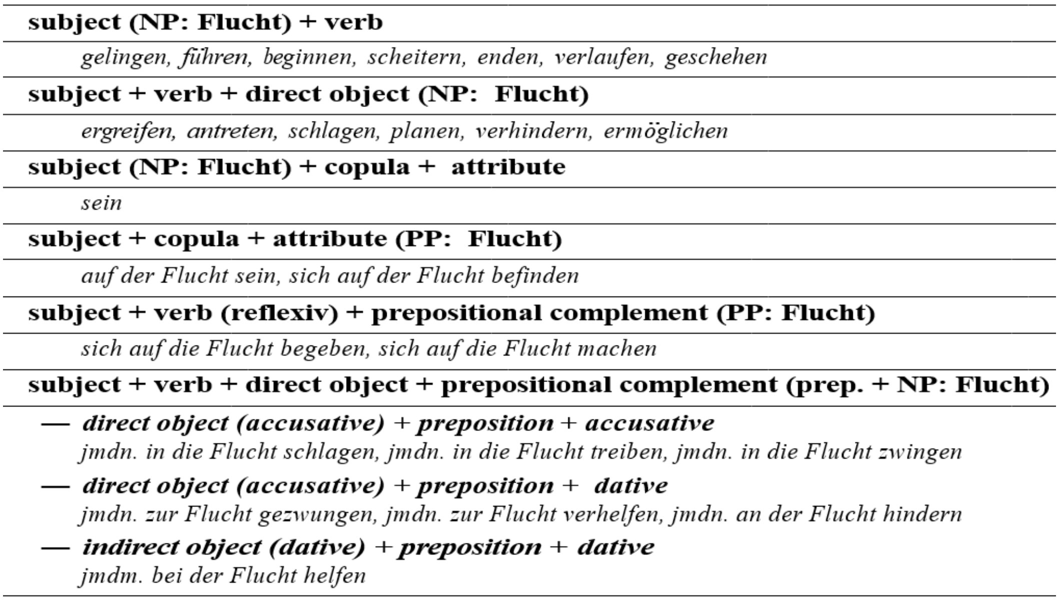
After this phase we will have to integrate the adjectives candidates to the Lexical Functions (LF) (Alonso Ramos, Tutin & Lapalme, 1995; Mel'čuk, 1996; Barrios Rodríguez, 2010) in the nominal phrase and generate the sentence context. For this purpose, the selection of LF is based on frequency criteria according to corpora data from Sketch Engine. Returning to the example of FLUCHT, we observe that this noun frequently appears combined with adjectives such as *überstürtzt* 'hastily', *dramatisch* 'dramatic', *heimlich* 'secret', *feige* 'cowardly', *missglückt* 'unsuccessful', *schleunig* 'rapid', etc. From this initial frecuency selection, the adjectival lexical items are allocated to classes according to the LF, for example as Magn-speed (*überstürtzt*, *schleunig*) and Antibon (*feige*, *dramatisch*, *missglückt*), and then we debug and package for each LF[20]. In this way, we get more natural examples of the nominal phrase:

Magn-speed: *Eine/Die/Jene/Jede [überstürtzte schleunige, ….] Flucht*

Antibon: *Eine/Die/Jene/Jede [feige, dramatische, missglückte, ….] Flucht*

In the next step we focus on the selection of verbs for each of the central structures (see Table 4). We follow the same procedure as before. In this way we generate sentence contexts with the examples which represent the most frequent valency patterns.

[20] Evidently, 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.



subject (NP: Flucht) + verb
gelingen, führen, beginnen, scheitern, enden, verlaufen, geschehen
subject + verb + direct object (NP: Flucht)
ergreifen, antreten, schlagen, planen, verhindern, ermöglichen
subject (NP: Flucht) + copula + attribute
sein
subject + copula + attribute (PP: Flucht)
auf der Flucht sein, sich auf der Flucht befinden
subject + verb (reflexiv) + prepositional complement (PP: Flucht)
sich auf die Flucht begeben, sich auf die Flucht machen
subject + verb + direct object + prepositional complement (prep. + NP: Flucht)
— ***direct object (accusative) + preposition + accusative*** *jmdn. in die Flucht schlagen, jmdn. in die Flucht treiben, jmdn. in die Flucht zwingen* — ***direct object (accusative) + preposition + dative*** *jmdn. zur Flucht gezwungen, jmdn. zur Flucht verhelfen, jmdn. an der Flucht hindern* — ***indirect object (dative) + preposition + dative*** *jmdm. bei der Flucht helfen*

Table 4: Sentence frame for the German noun FLUCHT[21].

Along with the debugging of the phrasal context generation and sentence context there is a combined testing and control phase. This is required because the occurrence of some type of LF might show restrictions concerning the presence of a semantic class of verbs or with some of their arguments or modifiers. For example, with a result from MultiGenera we can obtain the completely acceptable nominal phrase such as a). However, its use in a sentence frame such as b) would be unacceptable from a semantic and communicative point of view:

a) *die gelungene Flucht der Deserteure*
‘the successful escape of the deserters’

b) *Die gelungene Flucht der Deserteure war eine Katastrophe.*
‘The successful escape of the deserters was a catastrophe’

5. Conclusions

This paper deals with the different design and development stages of prototypes for the automatic generation of linguistic data, which can be directly applied to obtain examples that provide noun argument patterns at phrasal and sentence levels. We focus in particular on the description of the combined method for three languages (Spanish, German and French). The tools presented here make it easier to explore ontologies linked to wordnets and to automate lexical selection procedures in the slot-filling of nominal arguments in the three languages. The final prototype for generating noun phrases is provided with both packaged lexical files and API queries in WordNet

[21] NP: *Flucht* appears in a nominal phrase. PP: *Flucht* appears in a prepositional phrase.

following the semantic characteristics of the nominal arguments concerned. Certainly, deploying all these developments for the three languages has also been an added challenge due to its contrastive approach. The developments implemented for MultiGenera and MultiComb would not have been possible without the use of a series of tools that were not initially conceived for the generation of natural language. However, the outputs of both projects can also be used freely to improve these or other tools. In this way, the custom-made tools, the packaged lexical files and all the data concerning the combinatorial relations of nominal arguments and its restrictions could be especially useful for new developments.

6. Acknowledgements

This work has been developed within the framework of the projects "Generación multilingüe de estructuras argumentales del sustantivo y automatización de extracción de datos sintáctico- semánticos" - MultiGenera (BBVA Foundation grants for Scientific Research Teams, 2017) and "Generador multilingüe de estructuras argumentales del sustantivo con aplicación en la producción en lenguas extranjeras" - MultiComb (funded by FEDER / Ministry of Economy, Industry and Competitiveness - State Research Agency / Project FFI2017-82454-P).

7. References

Abel, A. & Meyer, C. (2013). The Dynamics Outside the Paper: user contributions to online dictionaries. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century: thinking outside the paper: proceedings of the eLex 2013 conference.* Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 179-94.

Alonso Ramos, M., Tutin, A. & Lapalme, G. (1995). Lexical functions of explanatory combinatorial dictionary for lexicalization in text generation. In P. St. Dizier & E. Viegas (eds.) *Computational Lexical Semantics, Studies in Natural Language Processing.* Cambridge: University Press. R´evision de la pr´esentation au Second Seminar on Computational Lexical Semantics, Toulouse, 1992, pp. 351-366.

Álvez, J., Atserias, J., Carrera, J., Climent, S., Laparra, E., Oliver, A. & Rigau, G. (2008). Complete and consistent annotation of wordnet using the top concept ontology. In N. Calzolari, K. Choukri, B. Maegaard, J. Ariani, J. Odijk, S. Piperidis & D. Tapias (eds.) *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC'08). Marrakech, Morocco: European Language Resources Association (ELRA).

Barrios Rodríguez, M. A. (2010). El dominio de las funciones léxicas en el marco de la teoría sentido-texto. *Estudios de Lingüística del español*, 30, pp. 1-477.

Bentivogli, L., Forner, P., Magnini, B., Pianta, E. (2004). Revising WordNet domains hierarchy: Semantics, coverage, and balancing. In *Proceedings of COL- ING Workshop on Multilingual Linguistic Resources*, MLR '04. Stroudsburg, PA, USA:

Association for Computational Linguistics, pp. 101-108.

Bond, F. & Foster, R. (2013). Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics*: ACL- 2013, Sofia.

Domínguez Vázquez, M. J. & Valcárcel Riveiro, C. (2019). 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.* De Gruyter: Berlin, pp. 135-158.

Domínguez Vázquez, M. J., Valcárcel Riveiro, C. & Lindemann, D. (2018). Multilingual generation of noun valency patterns for extracting syntactic-semantical knowledge from corpora (MultiGenera). In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.* Ljubljana: Ljubljana University Press, pp. 847-854.

Domínguez Vázquez, M. J. (2018). Was sind Valenzwörterbücher? Sprachwissenschaft, 43(3), pp. 309-342.

Engel, U. (1996). Semantische Relatoren. Ein Entwurf für künftige Valenzwörterbücher. In N. Weber (ed.) *Semantik, Lexikographie und Computeranwendung.* Tübingen: Max Niemeyer, pp. 223-236.

Engel, U. (2009). *Deutsche Grammatik – Neubearbeitung.* München: Iudicium.

Gouws, R. (2014). Towards bilingual dictionaries with Afrikaans and German as language pair. In M. J. Domínguez Vázquez, F. Mollica & M. Nied Curcio (eds.) *Zweisprachige Lexicographie zwischen Translation und Didaktik.* Berlin: De Gruyter, pp. 249-262.

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

González Agirre, A. & Rigau, G. (2013). Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual central repository. *Linguamática*, 5(1), pp. 13-28.

Izquierdo Beviá, R., Suárez Cueto, A. & Rigau Claramunt, G. (2007). Exploring the automatic selection of basic level concepts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing.* Shoumen, pp. 298-302.

Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In R. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography.* Berlin, Boston: de Gruyter, pp. 517-524.

Mel'čuk, I. (1996). Lexical functions: a tool for the description of lexical relations in a lexicon. In L. Wanner (ed.) *Lexical functions in lexicography and natural language processing.* Amsterdam: John Benjamins, pp. 37-102.

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

Melchior, L. (2014). Ansätze zu einer halbkollaborativen Lexikographie. *Online publizierte Arbeiten zur Linguistik*, 4, pp. 27-48.

Müller-Spitzer, C. (ed.) (2014). Using Online Dictionaries. Berlin/Boston: de Gruyter.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. J. (1990). Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4), pp. 235-244.

Niles, I. & Pease, A. (2001). Towards a standard upper ontology. In *FOIS '01. Proceedings of the International Conference on Formal Ontology in Information Systems*. New York: ACM, pp. 2-9.

Padró, L. (2011). Analizadores multilingües en freeling. *Linguamatica*, 3(2), pp. 13–20.

Sagot, B. & Fišer, D. (2008). Building a free French wordnet from multilingual resources. In N. Calzolari, K. Choukri, B. Maegaard, J. Ariani, J. Odijk, S. Piperidis & D. Tapias (eds.) *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC'08). Marrakech, Morocco: European Language Resources Association (ELRA).

Valcárcel Riveiro, C. & Domínguez Vázquez, M. J. (2016). Teste "muerte": falantes a avaliar a aceitabilidade de frases nominais geradas artificialmente. [https://carlosvalcarcel.net/2016/11/30/teste-muerte-falantes-a-avaliar-a-aceitabilidade-de-frases-nominais-geradas-artificialmente/]. Accessed 2 June 2019