# Automatic Lexicographic Content Creation for Lexicographers

María José Domínguez Vázquez[1], Daniel Bardanca Outeiriño[2], Alberto Simões[3]

[1] Universidade de Santiago de Compostela – ILG, Santiago de Compostela, Spain

[2] Universidade de Santiago de Compostela – Santiago de Compostela, Spain

[3] 2Ai, School of Technology, IPCA, Barcelos, Portugal

E-mail: majo.dominguez@usc.es, daniel.bardanca@rai.usc.es, asimoes@ipca.pt

**Abstract**

This paper presents *Combinatoria*, a tool for the semi-automatic generation of biargumental valency patterns for nominal phrases, as well as the current development of the tool for describing the passive valency of the noun. First, we describe a set of prototypes developed as exploratory tools for this new approach, together with the lexical and syntactic resources required for the generation of nominal phrases. We will focus especially on lexical resources, their automatic retrieval, and how they assist the lexicographic team in their tasks. This is followed by a description of the tool, the data filtering process, and the presentation of the obtained results. Finally, we include a brief discussion on the usefulness of these generators not only as stand-alone plurilingual dictionaries, but also as integrated resources in other electronic tools.

**Keywords:** multilingual valency dictionaries; argument patterns; automatic language generation; natural language processing

## 1. Introduction

The cooperation between lexicography and Natural Language Processing (NLP) has shown that the availability of lexical knowledge is beneficial at different levels (Trap-Jensen, 2018). This interaction, together with new developments in language technologies and the empowerment of the user of lexicographic resources, have significantly influenced the concept of the dictionary itself[1] and the types of tasks that lexicographers must undertake (Maldonado, 2019). According to Villa Vigoni-Theses (2018), the dictionaries of the future "are lexical or linguistic information systems in which existing lexicographic data are conflated, multilingualism and linguistic variety are entrenched [...]" and an essential task for lexicography is "the orderly conflation of data which has been generated automatically by text corpora and specifically processed [...]".

Regarding this context and considering the lack of resources for describing and consulting the valency of a noun, three prototypes for automatic generation of valency

[1] To understand this typological evolution, see, for example, Engelberg & Müller-Spitzer (2013) or Boelhouwer et al. (2017).

patterns were developed to show a new concept of electronic multilingual dictionaries, in this case, automatic and more interactive valency dictionaries (Prinsloo et al., 2011). The three simulators – *Xera*, *XeraWord*, and *Combinatoria* – have been designed as independent lexicographic tools for humans, but may also be integrated into other types of resources and even exported as computational lexicons (see Section 4). The main goal is to create a multilingual platform for describing and consulting the valency of different word classes. These generators also provide an innovative methodological approach: on the one hand, they combine different linguistic theories. – such as Valency Grammar, Prototype Theory, etc. – On the other hand, they implement NLP techniques, WordNet, Wordnet-like lexical databases, and other human-made multilingual resources for automatically generating lexicographic content (see Sections 2 and 3).

This study focuses on the tool *Combinatoria* (2020), a new prototype for automatic generation of biargumental valency patterns for nominal phrases in Spanish, German and French such as "*der Tod der Mutter an Tuberkulose*", "*la muerte del padre de infarto*", or "*la mort du marié par Ébola*". *Combinatoria* is not a stand-alone product; it is closely related to i) the monoargumental simulator *Xera* (2020), whose contents are used as the basis for the generation of nominal phrases with two arguments in *Combinatoria*, and to ii) the monoargumental simulator *XeraWord* (2020) that enables the automatic creation of examples for valency dictionaries in Galician and Portuguese. Although *XeraWord* and *Combinatoria* deal with the description of different languages, the first-mentioned tool allows us to analyse the feasibility of the data access structure – based on onomasiological criteria – and to implement it in the tool *Combinatoria*. The three generators, therefore, feed on each other; not only in terms of description levels and type of linguistic data fed to them, but also share applied analysis procedures and tools (Domínguez et al., 2019). They are free and are updated constantly.

While describing the tool *Combinatoria*, we highlight the role of the applied resources, in particular the set of tools we have developed for the automatic collection and generation of lexicographic content at different stages, as well as the work of the lexicography team (Jakubíček, 2018). Different human tasks are performed to ensure the quality of the automatically gathered data and check their accuracy regarding the dictionary type before being integrated into the generators. This study shows, therefore, how some automation procedures speed up lexicographic work and allow researchers to quickly adapt and design resources.

The paper is organised as follows. Section 2 focuses on the general features of the three language generators – *Xera*, *XeraWord*, and *Combinatoria* – including their description levels as well as the tools and procedures for their development. Section 3 deals with the current state of the project and future work. Section 4 presents the user interface, together with the user's data filtering process and the output of *Combinatoria*. Section 5 suggests possible further applications of this tool in the field of lexicography.

## 2. The language generators *Xera*, *XeraWord*, and *Combinatoria*

In this section we discuss the three tools that have been developed for the automatic generation of nominal phrases. First, a general description is provided. This is followed up by an explanation of the different procedures implemented during the development of the generators.

### 2.1. General description

The three generators provide information on the slots opened by a nominal head, that is, the active noun valency. Therefore, a specific slot for a given lexical unit is described considering its syntactic-semantic interface, as well as its combining potential and syntactic-semantic preferences (Engel, 1996; 2004). In opposition to other automatic language generators (Domínguez, 2020), the final goal of the tools is to answer the question of whether a noun *A* contains in its pattern an argument *X*, what their surface realisations are, and how each of them correlates with specific semantic-ontological classes and lexical units. This is the aim of *Xera* and *XeraWord*.

| | ***Xera*** | ***XeraWord*** | ***Combinatoria*** | ***CombiContext*** |
|---|---|---|---|---|
| language | es., fr., de. | gl., pt. | es., fr., de. | es., fr., de. |
| noun valency | active | active | active | passive |
| nouns | 60 | 10 | 60 | 60 |
| patterns | argumental | monoargumental | biargumental with phrasal context | ⟹phrasal and sentence context |
| chronology | first | third | version[1]: second version[2]: fourth | in progress |
| data access | formal: patterns | conceptual | conceptual | in progress |
| released | ✓ | ✓ | ✓ | - |

Table 1: General description of the generators

The focus is also on the combinatory potential, i.e., describing whether an argument *X* can be combined with another argument *Y,* and what restrictions or preferences determine this combination of arguments. The tool *Combinatoria* can provide this kind of information. It enables the user to obtain examples according to different surface realisations, after selecting the specific semantic role and semantic classes[2]. A new tool is already under development – *CombiContext* – for describing the passive valency of the nominal phrase, which will display its relationship to other units higher in the

[2] This relational and ontological approach differentiates *Combinatoria* from databases and annotated corpora such as CPA, Framenet, PropBank or Verbnet.

dependency hierarchy.

Table 1 summarises the general characteristics of the designed generators. As the starting point for verifying the feasibility of the methodological proposal and, ultimately, the prototypes themselves, 20 nouns in each language have been selected as representatives of different cognitive scenes or semantic fields (Table 2).

| | |
|---|---|
| MOVEMENT | huida-Flucht-fuite ‖ viaje-Reise-voyage ‖ mudanza-Umzug-déménagement |
| LOCATION | presencia-Anwesenheit-présence ‖ ausencia-Abwesenheit-absence ‖ estancia-Aufenthalt- séjour |
| EXPRESSION | conversación-Gespräch-conversation ‖ discusión-Diskussion-discussion ‖ pregunta-Frage-question ‖ respuesta-Antwort-réponse ‖ texto-Text-texte ‖ video-Video- vidéo |
| AFFECTION | muerte-Tod-mort ‖ aumento-Zunahme-augmentation ‖ dolor-Schmerz-douleur ‖ amor-Liebe-amour |
| CLASSIFICATION | olor–Geruch-odeur ‖ sabor-Geschmack-saveur ‖ color-Farbe-couleur ‖ (el) ancho-Breite-largeur |

Table 2: Nouns selected for generation

The descriptive levels for analysing the combinatory potential and rules of a language unit are common to the three currently available generators (Table 3).

| | **active valency** | | | **passive valency** |
|---|---|---|---|---|
| | ***Xera*** | ***XeraWord*** | ***Combinatoria*** | ***CombiContext*** |
| Only specific arguments are included in the argument pattern | ✓ | ✓ | +/- | +/- |
| Semantic description of the arguments: semantic roles | ✓ | ✓ | ✓ | ✓ |
| Semantic description of the arguments: ontological features | ✓ | ✓ | ✓ | ✓ |
| Syntactic function | ✓ | ✓ | ✓ | ✓ |
| Surface realisation | ✓ | ✓ | ✓ | ✓ |
| Interaction inside the nominal phrase | ✓ | ✓ | ✓ | ✓ |
| Interaction outside the nominal phrase | - | - | - | ✓ |

Table 3: Descriptions levels of the generators

A concrete example of these levels with some quantitative information is shown in Table 4 for the German noun *Diskussion* (*discussion*).

| | | |
|---|---|---|
| **Lemma** | • *Diskussion*<br>– Definition and semantic field<br>– Gender<br>– Number | |
| **Quantitative** | • monoargumental patterns: 23<br>• biargumental patterns: 78<br>• lexical packages: 111 | |
| **Syntactic-semantic** | • determinant+{adjective}+head+*über*+determinant argument<br>• determinant+{adjective}+head+*zwischen*+determinat+argument$_1$ *über*+determinant+argument$_2$ | |
| **Semantic** | • Relational | • Semantic role<br>– Role$_1$: someone, who discusses<br>– Role$_2$: what is being discussed |
| | • Ontological | • Ontological features<br>– Role$_1$: [animate ] [human]<br>– Role$_2$: [content] [situation] |
| **Morphosyntactic** | • Syntactic function<br>• Surface realisation | • subject /object<br>• *über* / *zwischen*+determinant + noun |

Table 4: Example of the information provided in the description

Since the properties of the nominal predicate determine the paradigm of lexical candidates that fit into a valency slot, getting and collecting these paradigmatic lexical units – or the *classe d'objets* according to Gross (2008: 11) – is key for subsequent programming. For the compilation of the lexical packages (see Section 3), it is necessary to consider that this vocabulary list must be filtered in such a way that it corresponds to the lexical units which fit into each of the argument slots of each argument for every surface realisation. Therefore, it is necessary to get and prototype a list of adequate lexical units[3] and encode their combinatorial rules and restrictions. This is dealt with in the next section.

### 2.2. Tools and procedures for developing the generators

This section provides a general overview of the common procedures applied as well as the tools developed or used to support the generators, relieve the workload of the lexicography team, and speed up the data compilation and revision procedures.

[3] To analyse and describe the syntactic-semantic interface we resort therefore to concepts such as semantic roles, ontological features, prototypical lexical units, and semantic classes (Domínguez et al., 2019; Domínguez, 2021).

Examples of some automation procedures will be presented in Section 3.

The steps and tools applied for developing the generators are summarised below: 1) setting the argument patterns: morphosyntactic and semantic analysis (Table 5), 2) Expansion and translation of lexical data (Table 6), 3) pre-integration into the generators (Table 7), 4) the generators themselves (see Section 4 for an example).

| **I. Setting the argument patterns: morphosyntactic and semantic analysis** | | | | | |
|---|---|---|---|---|---|
| **Goal** | **Collected data** | **Tools**[1] | | | **Human intervention** |
| | | **External available** | **Own created** | **Open access** | |
| Collecting the data | Frequency and valency data | | PORTLEX | - | Observation and data compilation |
| | | Sketch Engine | | - | |
| Establishment<br>• of the argument patterns: morphosyntactic | Morphosyntactic patterns | | PORTLEX | - | Data analysis and compilation according to valency criteria |
| | | Sketch Engine | | - | |
| • of the semantic roles: relational meaning | Patterns with semantic roles | valency dictionaries | | + | Lexical prototyping: development of an ontology and annotation of semantic classes |
| • of the ontological meaning | Patterns with ontological features | | bottom- up ontology | + | |

Table 5: Procedures and tools to establish argument patterns

An example of human intervention at this stage is the handling of data provided by Sketch Engine for the German noun *Diskussion* combined with a genitive case[4]. The corpora output cannot be automatically incorporated into the generators because: a) despite its high frequency, some surface realisations do not perform the function of a valency complement – for example, "Diskussion des letzten Jahrs" (*discussion of the last year*) or "Diskussion der letzten Woche" (*discussion of the last week*); b) the genitive of the noun "Diskussion" may express both those who discuss and the topic that is being discussed – for example, "Diskussion der Teilnehmer" (*discussion of the participants*) or "Diskussion der Ergebnisse" (*discussion of the results).*

This simple example illustrates that, in the first instance, frequency is not a crucial factor for selecting the lexical units that fit into a valency slot. In a second stage, frequency does indeed help us to determine lexical prototypes – lexical units that usually fit into a specific slot performing a well-defined semantic role. For example, the Argument$_2$ "what is being discussed" by [die Diskussion+determinant genitive+ Argument$_2$] in the meaning "die Diskussion einer Sache" (*discussion of something*) can be expressed with *Ergebnis (result)*, *Thema (tema)*, *Frage (question)*, *Begriff (concept)*,

[4] The CQL query was [lemma="Diskussion"][tag="(ART\.(Def|Indef)|PRO.(Dem|Poss).Attr).Gen.*"][tag="ADJ.*"]?[tag="N.*"].

*Problem (problem),* etc. We also analyse them according to general ontological features such as {content}, {situation}, etc. (for more information Domínguez, 2021; Domínguez et al., 2019). Once this is done, we are ready to undertake the next phase of the analysis: the expansion and translation of lexical data (Table 6). The aim here is to establish a controlled collection of a considerable number of lexical candidates.

| **Selection** | | | | | |
|---|---|---|---|---|---|
| **Goal** | **Obtained data** | **Tools** | | | **Human intervention** |
| | | **External available** | **Own created** | **Open acces** | |
| Selection | semantic relations of WordNet and ontologies linked to the synsets in the EuroWordNet model<br><br>Synset/meaning | WordNet | | + | Observation |
| | | | APIs | + | Tool development |
| | | | Lematiza | + | Selection of the synset and of the Wordnet ontological classe, with which an argument of the selected argument pattern fits.<br><br>**Benefit**: reduction of time spent in queries with a semi-automatic query selection. |
| **Expansion of the prototypes resorting to Wordnet** | | | | | |
| Getting new lexical units | Lexical collection of candidates, which share their characteristics with those of the lexical prototype. | | Combina | + | Tools development<br>Queries formulation and lexikal selection regarding the semantic classes, prototype and valency argument. |
| **Translation of the lexical unit's collection** | | | | | |
| Translation of lexical units | Lexical collection of candidates in other languages | | TraduWord | + | Tool development<br>Checking the translation quality<br><br>**Benefit**: to speed up the creation of new lexical packages por one language or to create new generators |

Table 6: Procedures and tools to get and compile new lexical candidates for different languages

In the generators, a valency-based description of the combinatory potential of the noun with a focus on the combinatory meaning (Engel, 2004) is of indispensable value. The question here is not only to find out whether a particular ontological entity fits into a valency slot performing a semantic role, but also which concrete lexical candidates or ontological features fit into it. The expansion procedure should not be underestimated, because diverse automatically generated data is key not only when using the resource, but also for its analysis from a qualitative point of view (Hashimoto et al., 2019; Vicente et al., 2015).

Once we have collected the lexical units that meet the requirements for being integrated into the generators, the steps described in Table 7 below are taken.

| **III. Pre-integration into the generators** | | | | | |
|---|---|---|---|---|---|
| **Goal** | **Obtained data** | **Tools** | | | **Human intervention** |
| | | **External available** | **Own created** | **Open access** | |
| Inflection | Inflected lemmas | FreeLing's dictionaries | | + | Checking the output |
| | | | Flexiona | + | Tool development |
| Paradigmatic packaging | Lexical packages | | | | Annotation to establish the descriptive levels required for the proper functioning of the generators |
| | Edited data | | Editor | - | Checking and correction |
| | New created lexical package | | Creador | - | Creation of lexical packages with paradigmatic information |

Table 7: Pre-integration procedures and tools

Due to the granularity of the linguistic levels (see Section 2.1; Table 3), the biargumental tool Combinatoria (see Section 4) leads to a total of 9,176 syntactic-semantic argument patterns for the nouns in Spanish, German and French[5], which implies an average of 152 combined structures per noun[6], for example:

- ['determinant', 'adjective', 'head', 'determinant genitive', 'argument N1G: {human political ideology}', 'über', 'determinant accusative', 'argument N3A: {intellectual meaning}']. Ex: *die alte Diskussion mit dem Faschisten über den Begriff.*
- ['determinant', 'adjective', 'argument N3:{intellectual meaning}', 'head', 'zwischen', 'determinant dative', 'argument N1D: {collective, group}']. Ex: *die rege Definitionsdiskussion zwischen den Delegationen.*

*Combinatoria* relies on *Xera*, which currently has the following analysed data[7] (Figure 1).

[5] In order to improve the semantic relevance of the combined structures, FastText models (Bojanowski et al., 2017) were also implemented for each language.

[6] Data on April 5, 2021.

[7] Examples for syntactic argument pattern order megastructure are [determinant+ adjective+Diskussion+über+argument N3A], [determinant+adjective+argument N3+ Diskussion], etc. Examples for syntactic-semantic argument pattern or interface syntactic-semantic are [determinant+adjective+Diskussion+über+argument N3A: {intellectual content}], [determinant+adjective+Diskussion+über+argument N3A: {intellectual meaning}], etc. Among the lexical units, the lemmas - for example *decano* (*Dean*) - from forms such as *decano*, *decana*, *decanos*, *decanas* (*Dean*, *Deans*) are differentiated.

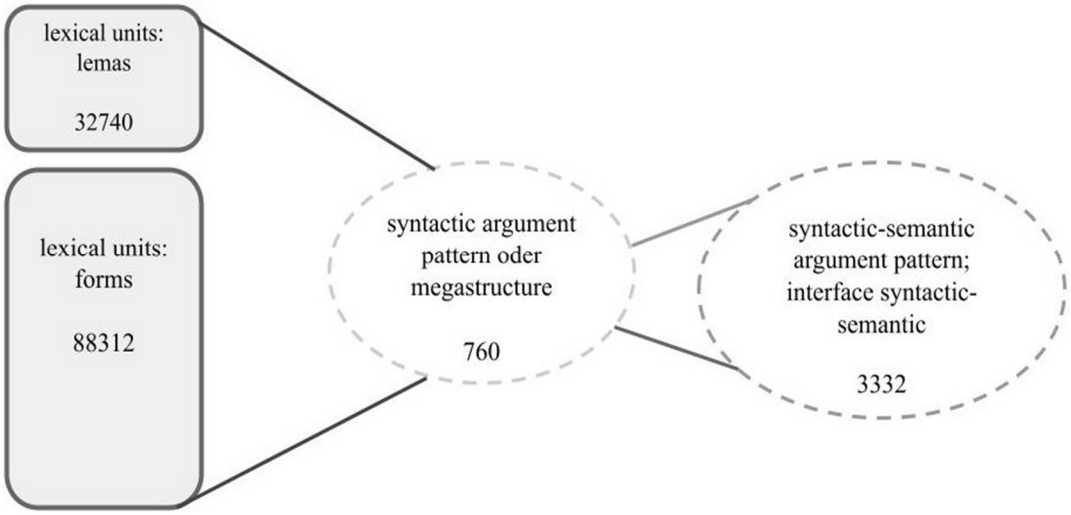


Figure 1: Current analysed data as the basis for *Combinatoria*

## 3. Resources

The tools presented here require a set of linguistic information that describes the different lexical units used in the phrases together with its semantic information, as well as the structure at sentence level for the integration of these units in the possible nominal phrases.

Although we are dividing this section into two parts, one for the description of the lexical resources and a second for the description of the syntactic structures, it is crucial to have in mind that these resources have coupled semantic information, as we will discuss.

### 3.1. Lexical Resources

The lexical resources used in *Xera*, *XeraWord,* and *Combinatoria* are structured following WordNet senses (in fact, its *synsets*), and based on a custom-tailored ontology derived from WordNet ontologies (see Section 2.2; Domínguez, 2020; 2021). This approach makes it possible to create a variety of phrases with the same or similar concepts, but compiling different words to guarantee the semantic validity of the generated sentences. This aids in the process of bootstrapping data for other languages.

A lexical package (see Table 7) describes a set of related lexical units that, although not interchangeable, have a similar paradigmatic relationship. As an extremely simple example, despite their different meaning, distinct parts of the human body can be used in similar structures – "the pain in my finger" or "the pain in my head" reproduce different meanings but share a common structure.

Each one of these lexical packages includes, for each valency slot of a noun, a unique identifier, a description of the type of object that is being characterised, its classification

in the ontology, and a list of lemmas. For each lemma, we link the respective Interlinguistic Index (ILI), used both in WordNet and the Multilingual Central Repository (MCR)[8].

*Xera* (see Section 2) started with three different languages: Spanish, German, and French. MCR does not include the French and German languages, but their wordnets were imported into the same database, and aligned using their ILI. This process allowed the creation of the original packages.

More recently, the Galician and Portuguese languages have also been included. In order to bootstrap the implementation of new languages, a set of tools were developed that help automate the translation by using WordNet and online translation services. These tools were first implemented in the development of *TraduWord* (see Table 6), which served to validate automatic translations of existing lexical packages and, therefore, to create automatic lexicographic content for lexicographers. The successful implementation of automatic translation of data circumvented the necessity to resource to raw WordNet data and subsequent debugging for every language (Domínguez et al., forthcoming). A concrete example of the implementation of *TraduWord* is the pilot tool *XeraWord,* which supports the Galician and Portuguese languages (see Section 2).

These lexical packages, while being the heart of *Combinatoria*, are useful in other contexts. Therefore, they are being codified using open standards and will be made available, independently of the online tool, in a public GIT repository.

### 3.2. Syntactic Resources

In the current stage of the project, we are developing a sentence generator, retroactively fed by all the previous work on simple and combined noun phrases. To successfully implement verb generation several previous steps were necessary.

So far, the focus was on semantically filtering appropriate nouns for the combination of noun phrases. At this point, there was no verbal data available in the database of the project. To supply this information, we developed resources based on open-source projects, namely a text chunker and a PoS (Part of Speech) tagger[9] that will allow the extraction of the relevant verbs, adverbs, and adjectives related to the so-called core nouns (Table 3). In this case, all data was extracted from Wikipedia text-only dumps.

Before starting the linguistic analysis of texts from these dumps, the original XML was preprocessed. In this case, the entry per se is the only relevant text we want to feed the NLP tools. Once this has been extracted, the results are stored in a spreadsheet with two columns. This allows us to keep track of the origin of each text (column 2) by

[8] Available at http://adimen.si.ehu.es/web/MCR.

[9] The parser and tagger used are part of the NLP library Spacy: https://spacy.io/

linking it to the headword used by Wikipedia (column 1). An extract from the data is shown in Figure 2.

| Standard | Standard | Standard |
|---|---|---|
| | headword | long_entry |
| 9 | Algorithmique | Algorithmique↵↵Lalgorithmique est l'étude et la production de règles et techniques qui s |
| 1 | Autriche | Autriche↵↵L'Autriche ( ), en forme longue la république d'Autriche (), est un État fédé |
| 2 | Algorithme | Algorithme↵↵Un algorithme est une suite finie et non ambiguë d'opérations ou d'instruct |
| 3 | Afghanistan | Afghanistan↵↵L'Afghanistan, en forme longue la république islamique d'Afghanistan (pacht |
| 4 | Auvergne | Auvergne↵↵LAuvergne ("Auvèrnha" en occitan') est une région culturelle et historique de |
| 5 | Alpes-de-Haute-Provence | Alpes-de-Haute-Provence↵↵Les Alpes-de-Haute-Provence ou AHP ( ), appelées Basses-Alpes |
| 6 | Alpes-Maritimes | Alpes-Maritimes↵↵Les Alpes-Maritimes ( ) sont un département français de la région Prove |
| 7 | Argentine | Argentine↵↵L'Argentine, en forme longue la République argentine, ( et "" ) est un pays d |
| 8 | Aka | Aka↵↵Aka peut désigner :↵↵Aka ou AKA peut désigner :↵↵Aka peut désigner :↵↵AKA peut fa |
| 9 | Aïkido | Aïkido↵↵L' est un art martial japonais (budo), fondé par Morihei Ueshiba "ōsensei" entre |
| 10 | Alliage | Alliage↵↵Un alliage est la combinaison d'un élément métallique avec un ou plusieurs méta |
| 11 | Arménie | Arménie↵↵L'Arménie, en forme longue la république d'Arménie, en arménien ', et ', , est |
| 12 | Angola | Angola↵↵L'Angola, en forme longue la république d'Angola, en portugais , en kikongo , es |
| 13 | Andorre | Andorre↵↵LAndorre, en forme longue la principauté d'Andorre (en catalan et ), est un Éta |
| 14 | Antigua-et-Barbuda | Antigua-et-Barbuda↵↵Antigua-et-Barbuda ou Antigue-et-Barbude est un État des Antilles a |
| 15 | Apple | Apple↵↵Apple ( « pomme » en anglais) est une entreprise multinationale américaine qui c |
| 16 | Astronomie | Astronomie↵↵L'astronomie est la science de l'observation des astres, cherchant à expliq |
| 17 | Abréviation | Abréviation↵↵Une abréviation (du latin "brevis", en français : « court », abrégé en « ab |
| 18 | Atoum | Atoum↵↵Atoum ou Toum (traduit par certains par "l'Indifférencié") est un dieu de la myt |
| 19 | Aton | Aton↵↵Aton est un dieu solaire de l'Égypte antique. Il est surtout connu comme un dieu é |

Figure 2: Spreadsheet with texts from Wikipedia

The PoS tagging pipeline is then applied to these sentences. The results are reorganised and stored in a tree-like structure that allows retrieval of the data by its frequency with the relevant noun as the central element. This enables the development of a user-oriented tool that lets researchers and language learners visualise the most common PoS tags at each position, together with the most common lemmas, always considering what has already been selected. Therefore, this approach allows autonomous development of new sentences by telling the machine what the desired output structure should have.

This data, together with previously developed work from the *Xera* and *Combinatoria* tools, are currently being used for the development of sentence-capable lexical packages. The new procedure takes up from where the original noun phrase combination phase left off, and the already combined noun phrases are further developed to include verbal constructions. Any modification of the original combined structure is possible with this tool, including the complete overhaul of the elements and data to make a new verbal combination. Four main elements are presented for immediate addition to the combined structures: adverb, verb, adjective, and nouns. Manual addition of other tags already supported by the system is also possible. These tags allow manual construction of combined verbal structures through fixed patterns. The information to fill in these tags can be chosen by following the user interface, as shown in Figure 3. New noun slots may be filled with lexical data from any previous ontological item already classified. A new module is being developed to process verbal combinations that will be called after the original *Combinatoria* module (Figure 3).

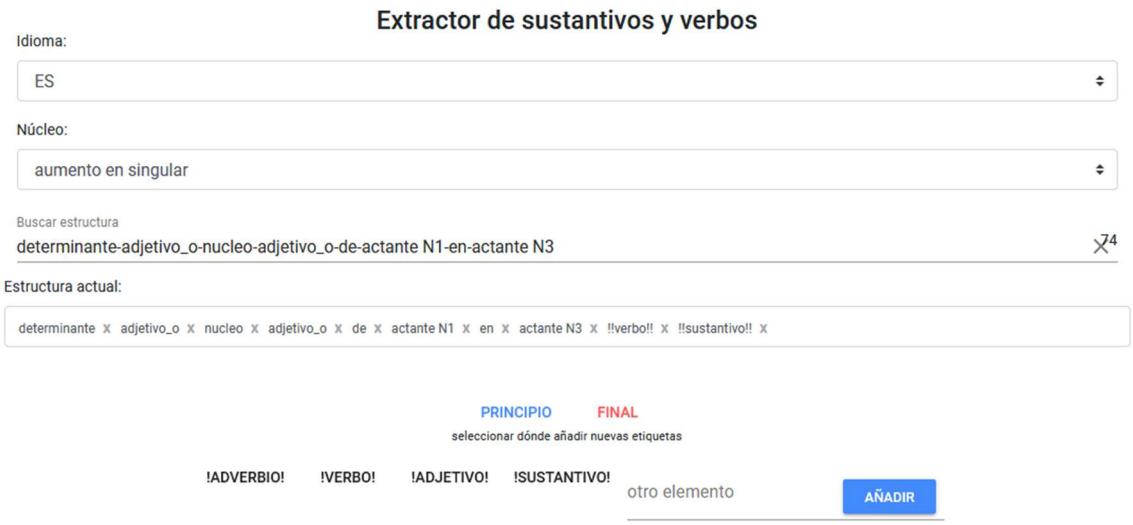


Figure 3: Verbal combinator

Tags marked inside exclamation marks will be processed as the last step, allowing the original *Xera* and *Combinatoria* projects to remain unaltered, while still being called to process already existing tags.

## 4. Using *Combinatoria*

When using the biargumental tool *Combinatoria*[10] (see Section 2.1), the user must first choose a target language and noun (Figure 4). The information about its meaning and semantic field is displayed as a mouseover effect.

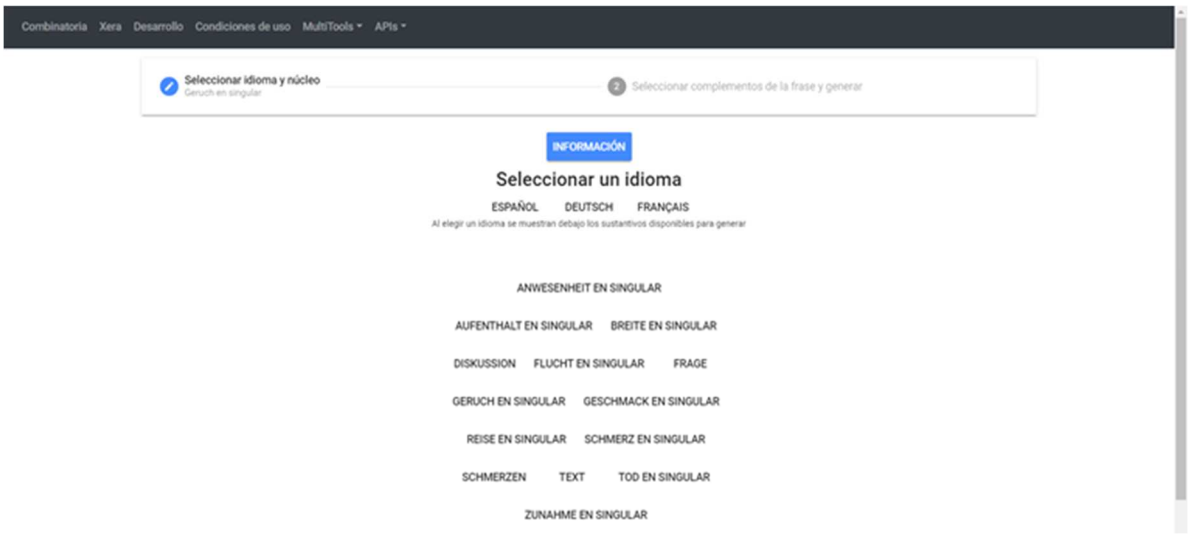


Figure 4: The main interface of *Combinatoira*

[10] Available at http://portlex.usc.gal/combinatoria/

Once the noun has been selected, e.g., *Geruch* (in singular) in German, the user decides which ontological-semantic feature should appear as the first argument. Let us suppose that the user selects *location – building – room* for the first argument, there are now two possible options:

a) As with argument 1, the user tunes the search options for argument 2 in the drop-down menu on the left (Figure 5):


Figure 5: Search options: ontological approach and filtering

b) A list is displayed in the middle of the screen containing all the possible combinations that align with the already selected filter (Figure 6). On the left, the user can see an example of what would be generated when selecting a specific item. The semantic classification for each argument that will be combined is displayed on the right side.

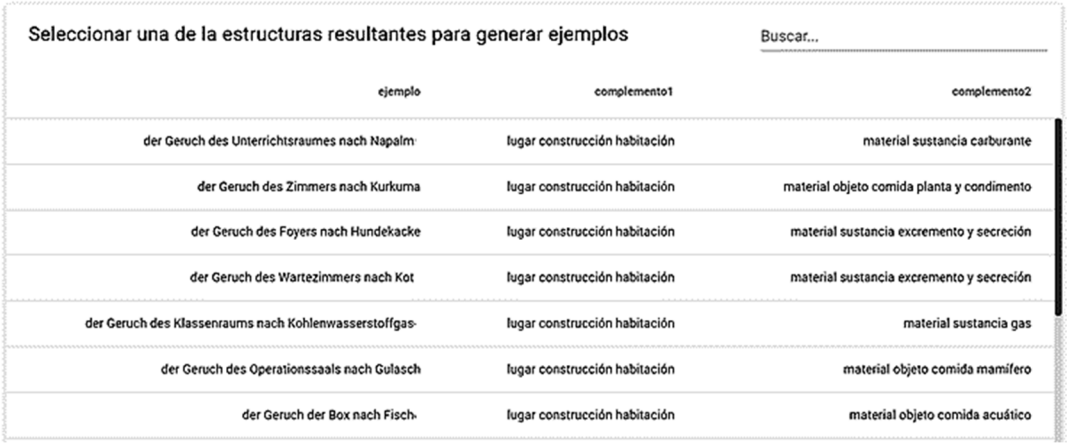


Figure 6: Search options with examples

Continuing with this hypothetical use of the tool, if the second argument is chosen as {*material - object - food - plant - condiment*}, the possible results which are combinations of the selected first and second arguments are shown (Figure 7):

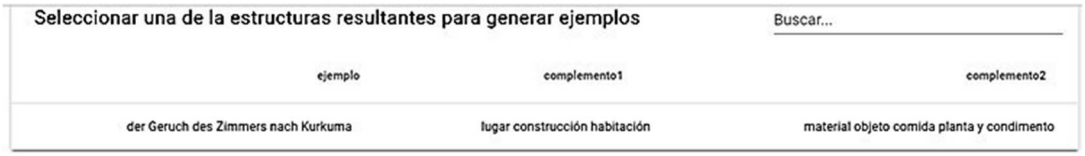


Figure 7: Structure selected for example generation

Upon clicking on one of the displayed possible combinations, the examples will be generated automatically (Figure 8). It is also worth adding that the generated data follows a principle of predetermined randomness. This randomness affects the lexical representatives of each class, but not the semantic role.


Figure 8: Automatic generated examples

The main novelty of the new *Combinatoria*, compared to its first version (Domínguez, 2020; Figure 9), is that it proposes conceptual onomasiological access to the argument pattern of the nominal phrases, as well as standard examples that can guide the user on the type of information that each label refers to. This approach avoids unnecessary valency terminology and formal abbreviations of roles and functions.

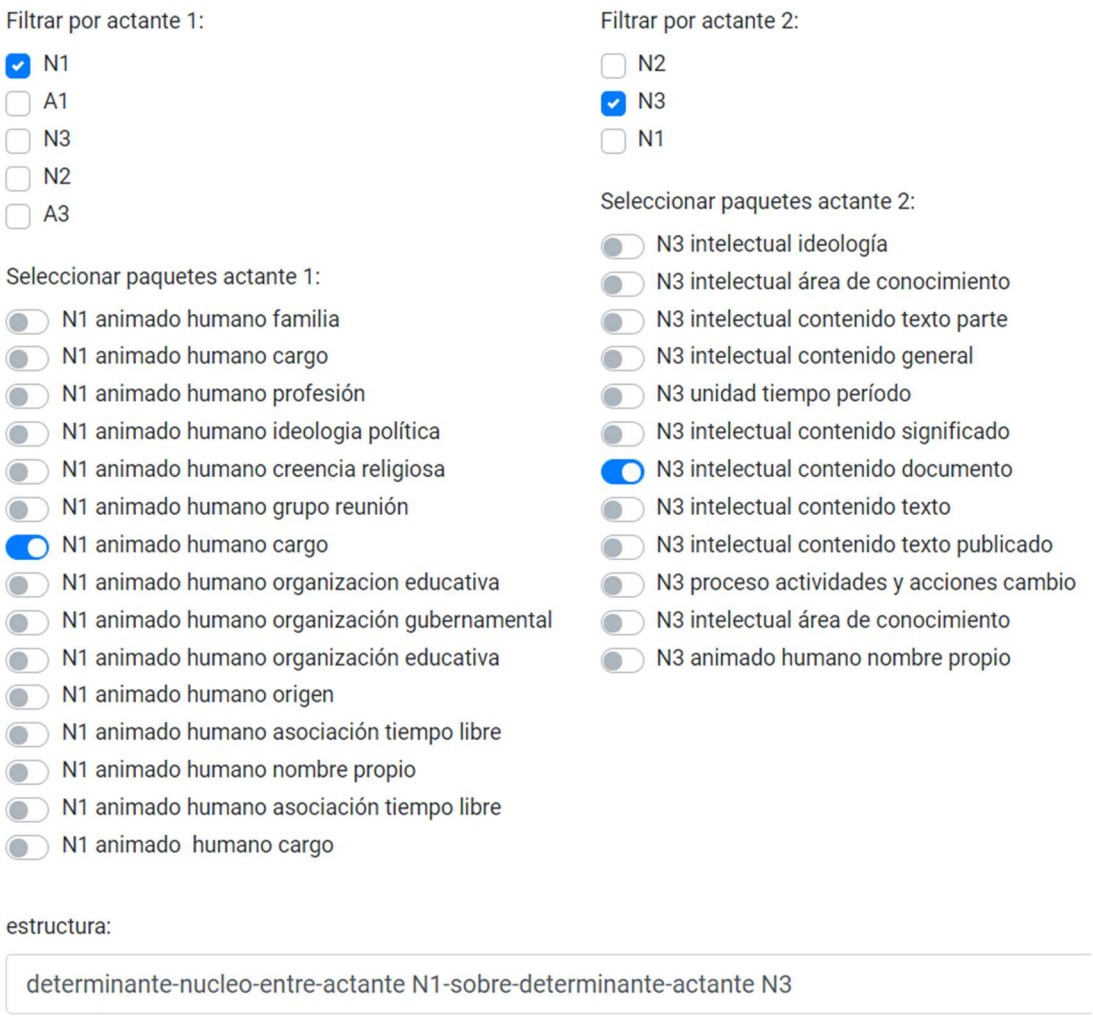


Figure 9: The user interface of Combinatoria 1.0

The primary users of our resources are foreign language learners and teachers. It should be highlighted here that the lexeme acquisition is bound up with the learning of its syntactic-semantic frame (Laufer & Nation, 2012) as well as that, in foreign language production, a considerable number of errors lie in the valency domain (Gao & Haitao, 2020; Nied, 2014; Müller-Spitzer et al., 2018).

Although we did not collect a scientifically representative amount of data on the use of these tools, some exploratory experiments with learners of German as a foreign language with A2-B1 level indicate that it takes time to understand the functioning of the tools *Xera* and *Combinatoria*. Taking into account the users' feedback, we are currently exploring the possibility of adding to the general information in the resources a step-by-step guide highlighting each required step. This will avoid unnecessary saturation of the user's interface with explanations and multiple choices. From these preliminary experiments, no preference for formal or conceptual access structure is concluded. Further studies among both learners and teachers are planned to better understand how users want to access the syntactic structures and how to improve the interface. This will also be done for the new *CombiContext* tool, described in Section 3.

## 5. *Combinatoria* for Lexicographic Work

As key applications of our tools (see Sections 2.2 and 3), and especially for *Combinatoria*, in the field of lexicography, we propose the following:

- As a stand-alone resource: primarily for lexicographic application, *Combinatoria* offers a verified methodological approach and serves as a prototype for further development of plurilingual valency dictionaries in other languages. To improve its usability, the number of units described in the system needs to increase in the future. The automation of analysis procedures, as well as the tools already designed (see Section 2.2) for the compilation and analysis of the lexical units and its semi-automatic translation (see Section 3) facilitates not only the integration of new languages but also the addition of lexical units for each of the prototypes. The first step in that direction has already taken place with the monoargumental tool for Galician and Portuguese *XeraWord* (see Section 3). To use the generators more efficiently in language teaching but also to develop lexicographic resources offering comparative information, it is possible to transform the generators into cross-lingual tools, similarly to the multilingual dictionary *Portlex* (2018).

- As an integrated resource into other dictionaries: It is worth highlighting the usability of the generators themselves as part of the dictionary's microstructure so that instead of static examples there would be dynamic examples, which could be selected by the user according to a specific query. Thus, the dictionary entry and the query itself are individualised.

From the point of view of the lexicographic team and their various tasks, the tools supporting the development of the generators (see Section 2.2) can streamline the human workflow for other projects on the syntactic-semantic interface, and especially in those resorting to WordNet.

## 6. Conclusions

A valency-based description of the combinatory potential of the noun with a focus on the combinatory meaning (Engel, 2004) is of indispensable value, especially for foreign language teaching and learning.

The question here is not only whether the particular ontological entity can (or cannot) fit into a valency slot in the rendering of a semantic role, but also which concrete lexical candidates or ontological categories can. This is the aim of the *Combinatoria* tool: to present a novel methodological approach for describing the noun valency. As valency resources themselves, the generators described in Section 2 are also innovative in that they enable an individualised selection of examples with specific ontological features as well as their generation *ad libitum*.

The integration of the generators into other lexicographic resources as well as their use as independent multilingual valency dictionaries require further automation of

collection and analysis procedures. It is also important to enlarge the scope of the tool by increasing the number of units and performing studies to improve the user interface.

## 7. Acknowledgments

This contribution has been developed within the framework of research "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, Spain), the research "Ferramentas TraduWord e XeraWord: tradución de caudal léxico e xeración automática da linguaxe natural en galego e portugués" (financed by the University of Santiago de Compostela as part of the programme "Convocatoria de proxectos colaborativos para institutos de investigación da USC. 2020"), and partly funded by Portuguese national funds (PIDDAC), through the FCT – Fundação para a Ciência e Tecnologia and FCT/MCTES under the scope of the project UIDB/05549/2020.

## 8. References

Boelhouwer, B., Dykstra, A. & Sijens, H. (2017). Dictionary Portals. In P. A. Fuertes-Olivera (ed.) *The Routledge Handbook of Lexicography.* London/New York: Routledge, pp. 754 –766.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, pp. 135–146.

Domínguez Vázquez, M. J., Simões, A., Bardanca Outeiriño, D., Caiña Hurtado, M. & Iglesias Allones, J. (forthcoming). Automatic Generation of Nominal Phrases for Portuguese and Galician Combining Multilingual Resources into XeraWord.

Domínguez Vázquez, M. J. (2021). Zur Darstellung eines mehrstufigen Prototypbegriffs in der multilingualen automatischen Sprachgenerierung: vom Korpus über word embeddings bis hin zum automatischen Wörterbuch. *Lexikos,* 31, pp. 1-31.

Domínguez Vázquez, M. J. (2020). Aplicación de WordNet e de word embeddings no desenvolvemento de prototipos para a xeración automática da lingua. *Linguamática*, 12(2), pp. 71-80.

Domínguez Vázquez, M. J. & Valcárcel Riveiro, C. (2020). PORTLEX as a multilingual and cross-lingual online dictionary. In M. J. Domínguez Vázquez, M. Mirazo Balsa & C. Valcárcel Rivero (eds.) *Studies on multilingual lexicography.* Berlin: de Gruyter, pp. 135-158.

Domínguez Vázquez, M. J., Solla Portela, M. A. & Valcárcel Riveiro, C. (2019). Resources interoperability: Exploiting lexicographic data to automatically generate dictionary examples. In I. Kosem & T. Zingano Kuhn (eds.) *Proceedings of the VI. eLex conference Electronic lexicography in the 21st century*: *Smart Lexicography.* Brno: Lexical Computing CZ s.r.o, pp. 51-71.

Engel, U. (2004). *Deutsche Grammatik – Neubearbeitung.* München: Iudicium.

Engel, U. (1996). Semantische Relatoren. Ein Entwurf für künftige Valenzwörterbücher.

In N. Weber (ed.) *Semantik, Lexikographie und Computeranwendung.* Tübingen: Niemeyer, pp. 223-236.

Engelberg, S. & Müller-Spitzer, C. (2013). Dictionary portals. In R. Gouws et al. (eds.): *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography.* Berlin: de Gruyter, pp. 1023-1035.

Fuertes Olivera, P., Niño Amo, M. & Sastre Ruano. A. (2019). Tecnología con fines lexicográficos: su aplicación en los Diccionarios Valladolid-Uva. *RILE. Revista Internacional de Lenguas Extranjeras*, 10, pp. 75-100.

Gao, J. & Haitao. L. (2020). Valency Dictionaries and Chinese Vocabulary Acquisition for Foreign Learners. *Lexikos*, 30, pp. 111-142.

Gross, G. (2008). *Les classes d'objets.* Paris: Presses de l'Ecole normale supérieure.

Hashimoto, T.B., Zhang, H. & Liang, P. (2019). Unifying human and statistical evaluation for natural language generation. In J. Burstein et al. (eds.) *Proceedings of the 2019 Conference of the North American Association for Computational Linguistics: Human Language Technologies.* Minneapolis: Association for Computational Linguistics, pp. 1689-1701.

Jakubíček, M. (2018). Practical Post-Editing Lexicography with Lexonomy and Sketch Engine. *XVIII EURALEX International Congress: Lexicography in Global Contexts.* Ljubljana. http://videolectures.net/euralex2018_jakubicek_sketch_engine/

Laufer, B. & Nation, P. (2012). Vocabulary. In S. M. Gass. Hrsg., Susan M. Gass & A. Mackey (eds.) *The Routledge Handbook of Second Language Acquisition.* London/New York: Routledge, pp. 163-176.

Maldonado, M. C. (2019). Las investigaciones de mercado en lexicografía comercial: un aprendizaje para el mundo académico e investigador. *RILE. Revista Internacional de Lenguas Extranjeras*, 10, pp. 101-118.

Müller-Spitzer, C., Domínguez Vázquez, M.J., Nied Curcio, M., Silva Dias, I. M. & Wolfer, S. (2018). Correct Hypotheses and Careful Reading Are Essential: Results of an Observational Study on Learners Using Online Language Resources. *Lexikos*, 28, pp. 287-315.

Nied, M. (2014). Die Benutzung von Smartphones im Fremdsprachenerwerb und -unterricht. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus.* Bolzano/Bozen: Institute for Specialised Communication and Multilingualism, pp. 263-280.

Prinsloo, D. J., Heid, U., Bothma, T. & Faaß, G. (2011). Interactive, dynamic electronic dictionaries for text production. In I. Kosem & K. Kosem (eds.) *Electronic lexicography in the 21st Century: New Applications for New Users, Bled.* Eslovenia: Trojina, Institute for Applied Slovene Studies, pp. 215-220.

Trap-Jensen, L. (2018). Lexicography between NLP and Linguistics: Aspects of Theory and Practice. 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. 25-37.

Vicente, M., Barros, C., Peregrino, F., Agulló, F. & Lloret, E. (2015). La generación

de lenguaje natural: análisis del estado actual. *Computación y Sistemas*, 19, pp. 721–756.

Villa Vigoni-Thesen, V. (2018). Dictionaries for the Future – The Future of Dictionaries. Challenges to Lexicography in a Digital Society. https://www.emlex.phil.fau.eu/files/2019/03/Villa-Vigoni-Theses-2018-English.pdf

**Dictionaries and tools**

CPA = http://www.pdev.org.uk/

Combina = http://portlex.usc.gal/develop/combina.php

*Combinatoria* (2020) = *Combinatoria. Prototipo online para la generación biargumental de la frase nominal en alemán, español y francés.* Universidade de Santiago de Compostela. http://portlex.usc.gal/combinatoria

Flexiona = http://portlex.usc.gal/develop/flexiona.php

Framenet = https://framenet.icsi.berkeley.edu/fndrupal/

FreeLing's dictionaries = http://nlp.lsi.upc.edu/freeling/node/1

Lematiza = http://portlex.usc.gal/develop/lematiza/

Portlex (2018) = *Portlex. Dicccionario multilingüe de la valencia del nombre.* Universidade de Santiago de Compostela. http://portlex.usc.gal/portlex/

PropBank = http://verbs.colorado.edu/propbank/framesets-english-aliases/

Sketch engine = https://www.sketchengine.eu

TraduWord = https://ilg.usc.gal/gl/proxectos/interoperabilidade-de-recursos-e-producion-automatica-de-linguaxe-natura

Xera (2020) = *Xera. Prototipo online para la generación automática monoargumental de la frase nominal en alemán, español y francés***.** http://portlex.usc.gal/combinatoria/usuario

XeraWord (2020) = *XeraWord. Prototipo online de xeración automática da argumentación da frase nominal en galego e portugués***.** http://ilg.usc.es/xeraword/en/

Verbnet = https://verbs.colorado.edu/~mpalmer/projects/verbnet.html

WordNet = https://wordnet.princeton.edu