Natural Language Processing (2025), **31**, pp. 606–630
doi:10.1017/nlp.2024.32

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Automatic generation of nominal phrases for Portuguese and Galician

MaríaJosé Domínguez Vázquez[1], Alberto Simões[2], Daniel Bardanca Outeiriño[3], María Caíña Hurtado[1] and JoséLuis Iglesias Allones[1]

[1]Universidade de Santiago de Compostela, Instituto da Lingua Galega - ILG, Santiago de Compostela, Spain, [2]2Ai, School of Technology, IPCA, Barcelos, Portugal, and [3]Universidade de Santiago de Compostela, CiTIUS, Santiago de Compostela, Spain
Corresponding author: María José Domínguez Vázquez; Email: majo.dominguez@usc.es

(Received 4 February 2023; revised 7 November 2023; accepted 12 February 2024; first published online 3 June 2024)

Special Issue on '**Natural Language Processing Applications for Low-Resource Languages**'

Abstract

This paper presents *XeraWord*, an innovative tool for automatically generating nominal phrases. *XeraWord* can be used for different tasks, ranging from teaching languages to the creation of examples in lexicography, or even for the development of resources for natural language processing. In this area, *Xera* was the first experiment, allowing the automatic generation of nominal phrases in three languages: German, French and Spanish. This tool was extended to support other languages, namely, Portuguese and Galician.

We start by presenting the theory behind the development of *Xera* and its new version, *XeraWord*, namely, the applied base methodology, and the natural language processing resources used to support it. Then, *TraduWord*, a tool specifically developed to construct resources for new languages, is presented. This tool allows the semi-automatic translation of the data required for the nominal phrase generation. For this, we discuss its advantages and disadvantages, analysing the quality of the translated resources, as well as the amount of manual work required to validate and correct these resources.

Keywords: natural language generation; WordNet; combinatory semantics; valency; ontologies

1. Introduction

Online resources on different word classes' valency and combinatorial frames are scarce. This statement as a starting point may appear surprising, if it is not considered that valency dictionaries and other combinatory resources, such as collocation dictionaries, are two different tool types. Both provide information about the combination of units, but they are based on two different concepts: 'collocation' and 'valency'. For example, consider a search in Sketch Engine for the structure *amor* [love] + adjective. The top results include examples that would be part of a dictionary of collocations, such as *amor verdadero* [true love], *amor incondicional* [unconditional love], *amor eterno* [eternal love] and *amor imposible* [impossible love]. However, none of these adjectives belongs to the valency pattern of the noun *amor*. In contrast to dictionaries of collocations (e.g. REDES (Bosque 2023)), valency dictionaries describe specific arguments, that is, those that make possible the complete production of the predicate in terms of grammatical and appropriateness of the predicate, for example, *amor materno* [maternal love].[a] Complement specificity

[a]Unlike the previous examples, the reformulation as *mother's love* or *the mother feels love* are possible here.

(Engel, 2004) is, therefore, the central feature in developing a valency dictionary. The scarcity mentioned above compromises the development of bilingual and multilingual resources based on valency patterns. Although there are some online valency verb dictionaries (such as *E-VALBU*[b] or *DCVVEA*[c]), there are no online resources available which present a complete framework for the combinatory potential of a nominal unit besides the monolingual dictionary *DICE*[d] and the multilingual *Portlex*.[e]

The starting point for developing tools for automatic language generation was the *Portlex* valency dictionary. *Portlex* is a multilingual, digital, semi-collaborative and cross-lingual resource which describes the nominal phrase taking into account its different arguments' surface realisations together with their combinatorial rules as well as their syntactic-semantic restrictions (Domínguez Vázquez and Valcárcel Riveiro 2020). The development of the multilingual dictionary *Portlex* led to the formulation of two multilingual pattern generators: *Xera* (Domínguez Vázquez 2020) and *Combinatoria* (Dominguez, Outeriño, and Simões 2021).[f]

These resources represent a new type of plurilingual valency dictionary that works as an information tool (Fuertes-Olivera, Niño Amo, and Sastre Ruano 2018, p. 79), which better meets the needs of the users through quick and simple queries that return individualised and concrete data (Spohr 2011). *Xera* generates simple nominal phrases with a mono-argumental structure. This tool can generate simple examples such as *el olor a flores* [the smell of flowers], and *Combinatoria* generates complex phrasal contexts as well as sentence contexts within the nominal phrases such as *el olor de la habitación a flores era agradable* [the smell of flowers in the room was pleasant]. Both tools provide the user with a complete and detailed description of the valency patterns of the nominal phrases in French, German, and Spanish (Domínguez Vázquez, 2020). Furthermore, the examples automatically generated by these tools may be used to develop new dictionaries or didactic resources. In contrast to other proposals for automatic language generation (Vicente *et al.* 2015; Roemmele 2016; Nallapati *et al.* 2016; Otter, Medina, and Kalita 2020) (see Section 2.2), *Xera* and *Combinatoria* are designed as dynamic valency resources with a didactic application (Prinsloo *et al.* 2011). They can be used by university students and teachers of foreign languages as interactive valency dictionaries.[g] We consider these tools as being dynamic and interactive because the user interacts with the resources by selecting options and therefore generating information *ad libitum*. In the future, the user will be able to generate syntactic-semantic patterns by coming up with a tentative individual query, that is, to create a specific example from scratch, and the resource will return whether this construction is already recorded in the resource or not. This option will be accompanied by the possibility to create exercises.[h]

Another prominent feature of these generators is their broad semantic foundation. To automatically generate nominal phrases and their arguments, they rely on semantic approaches such as semantic valency and combinatory meaning (Engel 2004), lexical functions (Mel'čuk, 1996) and lexical prototype theory. Furthermore, for the analysis and description of the syntactic–semantic interface, we also resort to concepts such as semantic roles, ontological features, prototypical lexical units, and semantic classes (see Sections 2.1 and 3.1).

Given that there are no tools for nominal phrase generation for Galician and Portuguese, we have decided to partially apply the already verified methodology of *Xera* (German, Spanish and French) to develop a new resource, *XeraWord*. Therefore, to reuse previous knowledge, we have

[b] *Elektronisches Valenzwörterbuch deutscher Verben*, https://grammis.ids-mannheim.de/verbvalenz

[c] *Diccionario de valencias español-alemán*, https://gramatica.usc.es/es/projects/15.

[d] *Diccionario de colocaciones del español*. http://www.dicesp.com/paginas.

[e] Available at http://portlex.usc.gal/.

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

[g] The availability of resources for describing valency and argument combinatory possibilities is especially important since various studies conclude that this is one of the fields in which foreign language students make many errors in production (Nied Curcio 2014; Gao and Liu 2020; Müller-Spitzer et al., 2018).

[h] For detailed information on the application of the tools in second language learning, see Domínguez Vázquez *et al.* (2019).

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

developed a tool for describing these two closely related languages, Galician and Portuguese, with a new methodological approach. Thus, we explain how a new tool is created for two minority languages, at least from the point of view of the existence of natural language generation resources.

The article is organised as follows. Section 2 explains the general theoretical framework for the development of the generators; in particular, we will discuss the valency and dependency approach and the generators' typology in their immediate context. Section 3 focuses on the applied methodology. Section 4 provides an extensive description of the resources used and the tools developed to process Galician and Portuguese. In Section 5 we explain how the user can use *XeraWord* and discuss some of the results that can be obtained. Lastly, the final section concludes on the project.

2. General framework for the development of the language generators

This section presents the general theoretical framework of the developed tools and briefly discusses different approaches to automatic language generation.

2.1 Valency and dependency grammar

The previously mentioned generators present the syntactic–semantic interface of the available valency nouns, enabling specific frames composed of fill-in slots (Engel 2004). We understand valency as a multidimensional concept that may be described as 'quantitative valency' (number of slots) or 'qualitative' (type of slots). These slots are also considered at different descriptive levels (syntactic, semantics or pragmatic). In the current phase, the generators show the active noun valency since a nominal head opens the described slots. An example of passive valency would be integrating these valency data within other frameworks, displaying their relationship to other units higher in the dependency hierarchy.

Portlex, *Xera* and *XeraWord* follow a conceptual approach to the valency grammar. In our model, the scenario (Fillmore 1977, p. 62) or network of scenes linked to the different concepts and meanings, which have been codified through different syntactic frames, plays a predominant and hierarchically superior role. Thus, a scenario of MOVEMENT covers different types of movements and therefore different mental representations (scenes) of them (Hermanns and Holly 2007; Ziem 2008; Busse 2012). Our approach assumes that speakers share common cognitive representations that cover the different situations or micro scenes (Smith 1991) and that, consequently, in a communicative act, a speaker activates different surface realisations in different situations, for example, a specific lexeme with its syntactic–semantic frame (Domínguez Vázquez, 2015).

To provide a linguistic description of an abstract concept such as 'scenario' and 'scene', it is not possible to resort to word classes (verbs, adjectives, etc.) or surface realisations: it is well-known that the same surface realisation may perform different syntactic roles, as well as functions. It is the presence or absence of a role (or combinations of roles) that allows the description of a scenario: this is what we call 'a focalised semantic role'. This relationship between scenario, scene and role is not random (Engel 1996, p. 227). Therefore, the presence of directive roles in a noun such as *viaxe* (trip) makes it possible to assign it to a MOVEMENT scenario. The same applies to other surface realisations such as verbs (*to travel*, *to go away*, *to run*) or even interjections (such as *out*). The fact that *viaxe* may update semantic roles of origin, transit and destination, explains the classification of this noun as movement–displacement (e.g. *the trip to Japan* and *the sea trip from Mallorca to Valencia*). Ultimately, to determine the existence of a 'displacement' scene, the implicit or explicit presence of a locative-directive semantic role is necessary, regardless of the surface realisation (see Figure 2).

The relational meaning (Engel 1996, 2004) determines the semantic relation of the complements of the nominal head. Due to the expansion of case inventories and the uncertainty in the

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

delimitation of some of the roles, different asymmetric proposals are observed. Polenz (2012) describes nineteen semantic roles, while Engel (2004) proposes four thematic roles. In semantically annotated corpora, there is also no confluence in the type and number of the semantic roles. For example, *Verbnet*[i] Kipper *et al.* (2000) consider up to twenty-three thematic semantic roles, which do not coincide with the semantic role label from *NomBank* (Meyers *et al.* 2004), which is based on the *PropBank* label set, nor with the roles proposed on *Framenet* (Ruppenhofer *et al.* 2006; Žabokrtský *et al.* 2020). Finally, Dowty (1991) and Foley and Valin (1984) advocate the establishment of two proto-roles (*proto-agent* and *proto-patient*) and two macro-roles (*actor* and *undergoer*). We define a set of underlying semantic roles based on the catalogue proposed by Engel (2004) and its multi-contrastive application (Domínguez Vázquez and Valcárcel Riveiro 2020). We contemplate four central roles: agent, affected, classificative and locative. To describe in more detail the granularity of the semantic and formal levels, these roles are indexed. For example, a locative may be stative (*He lives in Madrid*) or directive (*He comes from Madrid*) and may also refer to the origin, transit or destination of the movement. Based on this idea, a catalogue was developed following didactic guidelines (Domínguez Vázquez and Valcárcel Riveiro, 2020). These roles, unlike those used in *Portlex*, are not explicitly displayed to the user in the generators, but they are the theoretical ground that organises all realisations. This catalogue may also be applied as *tertium comparationis*, which allows the organisation and structuring of the linguistic data by the different languages described in the resources. Regarding the results of the semantic relation described in the resources, we agree with Zhu *et al.* (2019): 'Across different languages, semantic structures are much more uniform than syntactic structures. However, there are still language variations in shallow semantics'.

The properties of the predicate determine not only the semantic roles but also the paradigm of lexical candidates for expressing a specific role. They are the other side of the combinatorial meaning. Thus, 'the one who acts' in a noun such as 'discussion' cannot be a priori an object such as a stone. Therefore, it is necessary to get and prototype a list of adequate, prototypical lexical units or a specific lexicon (see Section 4). The ontological categories we used for prototyping are derived from our bottom-up ontological classification. This means that we categorise the lexical units once the vocabulary has been collected and after considering other ontological proposals (mainly WordNet (Miller *et al.* 1990). Our ontology contains prototypical lexical units for a specific slot to which we apply a lexical prototyping procedure, which is used to extract prototypical semantic classes (see Section 4).

Therefore, we describe the combinatory potential of a language unit: the specific slots for a given lexical unit according to the following rules:

- The noun is the predicate or head, and, therefore, we describe its active valency.
- The argument pattern or frame includes only specific arguments.
- The arguments can be described as attending to:
  - Semantic level: semantic roles and ontological features
  - Morphosyntactic level: syntactic function and surface realisation
  - Interaction in the nominal phrase that is handled by the combinatorial tool, *Combinatoria*

2.2 Automatic language generators

Xera (Domínguez Vázquez 2020; Domínguez Vázquez *et al.* 2019) and *XeraWord*[j] propose a new concept of automatic and interactive valency dictionary (Prinsloo *et al.* 2011). These tools

[i] Available at https://verbs.colorado.edu/~mpalmer/projects/verbnet.html.

[j] As no research work has yet been published on this tool, no bibliographical reference is provided.

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

have been designed as free, open-access and independent lexicographic tools for human access. Nevertheless, they can be used and integrated into other types of resources and can be used to export computational lexicons. Beyond the languages they describe, there is a remarkable difference between both; in fact, *Xera* generates formal data automatically, while *XeraWord* is a semantic-ontological tool. The second notable difference is that *XeraWord* is based on the automatic translation of vocabulary lists mapped to semantic classes and ontological features.

The main similarities or differences between our natural language generation tools and those available in their close area of influence are as follows:

- They offer argument patterns or examples related to the syntactic–semantic valency and the combinatorial potential of the noun. Consequently, they differ from other resources which generate different types of communicative units (Moreno Jiménez *et al.* 2020; Otter, Medina, and Kalita 2021) and also from others that provide automatic generation of lexicographic content (Bardanca Outeiriño 2020; Bovi and Navigli, 2017; Kabashi, 2018). They are also different from resources such as GDEX[k] (Kosem *et al.* 2018), because we provide a detailed valency description: it is not only about good examples but also those that are adequate for the target and topology of our resource.
- Our generators enable the user to select the information according to their surface realisation (*Xera*), the intersection between the surface level and the semantic classes associated with each specific role (*Combinatoria*) and according to ontological features (*XeraWord*). This differentiates them from databases and annotated corpora such as Verbnet,[l] Propbank,[m] CPA[n] or Framenet.[o]
- In the Natural Language Generation field, our simulators represent also a new model. Current text production robots, such as Inferkit[p] or Sassbook[q] (for more information, see the work by Køhler Simonsen (2020)), automatically create text from a reduced number of words as a starting point. These kinds of tools are based on corpora and they apply statistical information about sequences to formulate new sentences given a context. Unlike these, in our prototypes, lexical packages are semantically annotated and syntactic–semantic structures are tagged so that the generated sentences follow a syntactic–semantic pattern defined by the user. Therefore, the user decides practically all the steps of his query, almost as querying a corpus using the Corpus Query Language (CQL), with the exact type of noun phrase she wants to generate (see *XeraWord*, Figure 8) to the sentence context in which it should appear. This is not the case in other text generators, developed mostly for English (see e.g. Rytr[r]).
- Another important difference is the target users of our resources: students (in the first instance) and foreign language teachers. There is a need for accessible online resources that may be successfully applied in foreign language teaching. Alongside, the generators themselves, companion exercises and activities are being developed. The goal is also to achieve a new model for combinatorial dictionaries that focuses on the user of the digital resources, along with the individualisation of the requested content.

[k] *Good dictionary examples*, available at https://www.sketchengine.eu/guide/gdex/.
[l] Available at https://verbs.colorado.edu/~mpalmer/projects/verbnet.html.
[m] Available at http://verbs.colorado.edu/propbank/framesets-english-aliases/.
[n] Available at http://www.pdev.org.uk/.
[o] Available at https://framenet.icsi.berkeley.edu/fndrupal/.
[p] Available at https://inferkit.com/.
[q] Available at https://sassbook.com/.
[r] Available at https://rytr.me/.

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

Table 1. Corpora results for the search of *conversación con + determinant +director/-a* in Sketch Engine

Word	Frequency	Freq. per million	% of concordance
conversación con el director	223	0.0114	41.44981
conversaciones con el director	145	0.0074	26.95167
conversación con la directora	43	0.0022	7.99257
conversaciones con el Director	41	0.0021	7.62082
conversación com el Director	28	0.00143	5.20446
conversaciones con la directora	14	0.0007	4.83271
conversación con los directores	8	0.0004	1.48699
Conversación con la directora	4	0.0002	0.74349
Conversación con el Director	2	0.0001	0.37175
Conversaciones con el director	1	0.0001	0.18587

2.3 The automatically generated examples in Xera and XeraWord

Both tools generate argument patterns or examples of valency nouns. Examples are a core element in lexicography, with an extensive bibliography both on its function and typology (Jacinto García, 2015; Lettner, 2019). The authors describe several approaches to gather examples for lexicographic purposes. These examples can be obtained from real corpora (and presented to the reader with their origin and possible adaptation) or can be created manually, either made-up or 'controlled' examples (Humblé, 1998; Lettner, 2019). The lexicographer will choose the example taking into account (i) the addressee, (ii) the resource type (monolingual or bilingual, or, e.g. encyclopaedic), (iii) its conception as a production or reception tool and (iv) its purpose for illustrating or documenting words.

The most used approach is to cite examples obtained from corpora. They can be either included without modifications or considered as a model for lexicographical examples (Atkins and Rundell, 2008). Prinsloo *et al.* (2012), among others, stress that the examples found in corpora often do not meet the requirements of intelligibility or conciseness demanded for a particular dictionary. Laufer (1992) further indicates that the examples elaborated or treated by the lexicographer are pedagogically more beneficial than the authentic ones. (Rundell, 1998, P. 334) summarised it as: *there is a certain tension here between the desirability of showing authentic instances of language in use, and the need for examples that work as hard as possible for the user*.

Our tools adopt an intermediate approach: they are neither examples taken directly from the corpus, nor examples made up *ad hoc* by the lexicographic team. The decision not to handle examples directly drawn from corpora is due to the following aspects:

- We observe that the examples found in corpora are not always suitable for dictionaries. They have excessive context, pronouns and anaphora that hinder the proper display of the term usage, namely the argument pattern and the lexical candidates that can fill the valency slots. The difficulty in finding examples with an adequate vocabulary for every one of the possible combinations in five languages, and the lack of semantically annotated corpora, especially for some languages, such as Galician, cannot be underestimated.
- We could verify that corpora do not include the lexical variability that the generators do. For example, if we search in Sketch Engine for a noun phrase such as *la conversación con las directoras de empresa* [the conversation with the company directors (women!)], we do not find any hits. Table 1 shows that the masculine form *director* is more frequent than

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

the feminine *directora*. Also, the term never occurs in the plural form. This fact is rather a social reflex but not really a semantic restriction.

The *Xera* and *XeraWord* generators provide examples of grammatically correct and semantically acceptable noun phrases. They are not made up examples but neither examples extracted from corpora. They are based on corpora exploration, grammar and valency information and lexical database oriented, as will be made clear in the next section.

Our examples are intended to point out possible lexical combinations or syntactic-semantic restrictions of a valency unit (compare the work by (Jacinto García 2015, p. 28)), therefore fulfilling the function for which they have been created: to illustrate language use and not to document it (Lettner, 2019). These examples should help learners to infer grammatical rules and semantic constraints regarding specific realisations. Furthermore, generating customised examples could help teachers better refine the input they need to provide to their learners.

3. Methodology

The methodology applied in the development of the generator of nominal phrases in Galician and Portuguese is based on previous research, part of the *MultiGenera* and *MultiComb* projects[s] (Domínguez Vázquez *et al.* 2019). There are, however, new research proposals and tools developed purposefully for handling the Galician and Portuguese languages.

3.1 Core methodology

This section presents an overview of the methodology used by the generation tools, to contextualise the reader on the approach. Figure 1 represents the core methodology, combining concepts and techniques from different areas, and their application for the development of the original multilingual generators for Spanish, French and German.

This methodological approach consists of several analytical steps.

1. Argument patterns.
The syntactic-semantic argument pattern is the starting point. It is necessary to know the number of arguments or syntactic–semantic elements of each noun for a given sense. Considering the German noun, *Umzug* [move], it requires three specific arguments, as shown in Figure 2.

2. Prototyping: lexical and semantic classes.
To achieve the automatic generation of the arguments and their combinatory potential (regarding not only their grammatical correctness but also their semantic accuracy), it is necessary to analyse which lexical candidates can cover the paradigmatic axis of each argument (Section 4.3). It is required to know which lexemes may be represented semantically as *that which acts* (see Figure 2). To gather this information, a set of CQL queries, executed against Sketch Engine's[t] databases, are performed to understand which lexemes tend to appear in a specific functional slot. Following this first quantitative role assignment, a lexical prototyping process is carried out to assign concrete ontological features to the argument structures.

In the case of the noun *Umzug* (Figure 2), lexemes such as *Wohnung* [apartment], *Umgebung* [surrounding area] and *Haus* [house] appear frequently with a prepositional phrase with *aus* as the second argument (Argument 2). This lexical prototyping, based on our semantic-linguistic ontology, allows the accurate detection of which lexical prototypes may be used frequently to fill a specific argument *x*.

[s] Available at http://portlex.usc.gal/ and http://portlex.usc.gal/combinatoria/
[t] Available at https://www.sketchengine.eu.

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

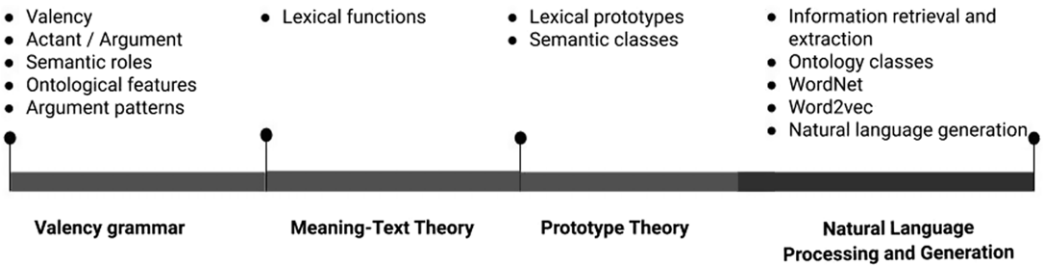

Figure 1. Methodological approach.

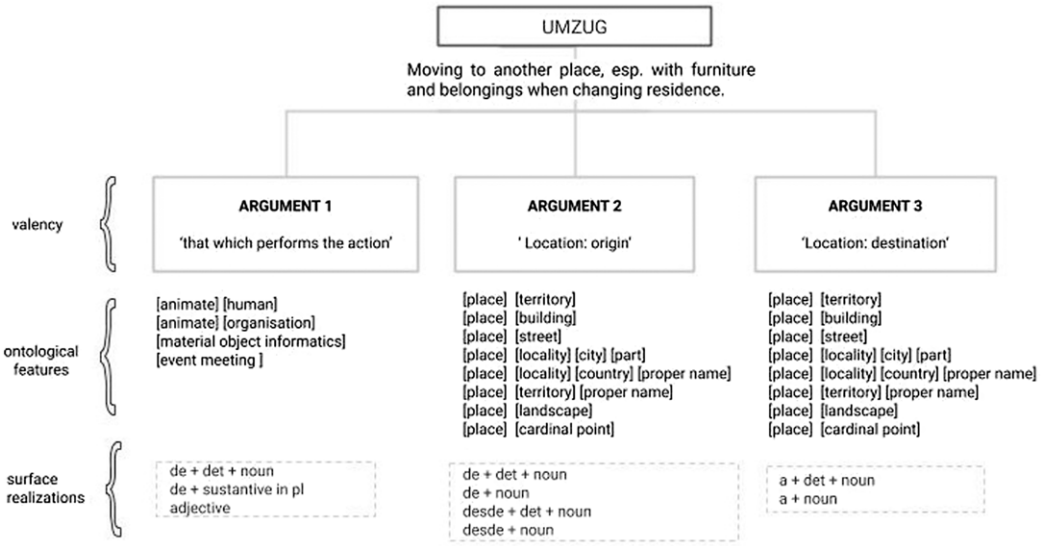

Figure 2. Semantic and formal description of the argument patterns for the German noun *Umzug*.

For example, for the Argument 3 preposition *aus*:

Wohnung, Gebäude, Haus, Villa: [+place] [+building]
Umgebung, Umfeld, Provinz: [+place] [+territory]
North, West: [+place] [+cardinal point]
Berlin, Rom: [+place] [+locality] [+city]

This analysis leads to a second step in the prototyping stage where semantic features are defined for each valency slot: for instance, {place building} or {place locality country proper name}. During these two steps, there is an intermediate step to filter irrelevant content.

3. Expansion of prototypes.

Both ontological features and lexical prototyping are key concepts that allow the easy semi-automatic extraction of lexical selections and future lexical expansion because they are fundamental for the linking of our classes to the semantic attributes present in WordNet (Gómez Guinovart, 2011; Gómez Guinovart and Solla Portela, 2018).

For this lexical expansion process (Figure 3), two tools were developed: a lemmatiser (*Lematiza*[u]) to compute all the lemmas with their different meanings (synsets) linked to the

[u] Available at http://portlex.usc.gal/develop/.

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

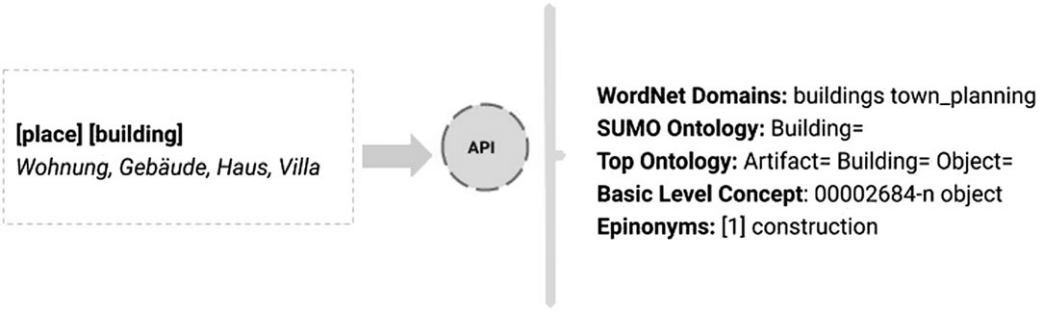

Figure 3. Lexical expansion using WordNet.

ontology of WordNet and the tool *Combina*[v] to compare the semi-automatic, shared or combined, results extracted from different queries with the following ontologies and knowledge bases—Top Concept Ontology (Álvez *et al.* 2008), WordNet Domains (Bentivogli *et al.* 2004), Suggested Upper Merged Ontology (Niles and Pease 2001), Basic Level Concepts (Izquierdo Beviá *et al.* 2007), Epinonyms (Gómez Guinovart and Solla Portela 2018) and semantic primitives (Miller *et al.* 1990).

This results in lexical input that shares the same semantic characteristics as the lexical-semantic prototype that was used as a starting point. This lexical selection produces a semantic class with a set of lexical candidates.[w]

4. Inflexion and paradigmatic packaging.
The produced lexical input needs to be inflected according to the morphological and semantic characteristics of each slot so that it can be implemented in the automatic generation tools as lexical packages (see Section 4.3).

5. Nominal phrase generation.
These lexical packages are then used for the generation of mono and bi-argumental structures. Two different tools are currently using this approach: *Xera* for mono-argumental patterns and *Combinatoria* for the bi-argumental patterns. In the case of *Combinatoria*, FastText models (Bojanowski *et al.* 2017) trained and published by Sketch Engine were implemented for each language to improve the semantic coherence of the results. We do not use word embeddings to disambiguate the meaning of tokens but to assess the contextual and semantic compatibility of both arguments in the generated examples. By comparing the cosine similarity of the vectors of each word present in the relevant lexical packages, we can filter out incompatible combinations. Embeddings are vectorial representations of words in an n-dimensional space. The cosine similarity is obtained by computing the cosine of the θ angle between these two vectors:

$$\frac{a^T b}{|a| \cdot |b|}$$

[v]Available at http://portlex.usc.gal/develop/combina.php.

[w]Both the list of lexical candidates extracted from WordNet and the list of units translated with TraduWord (see Section 4.5) contain complex phrases, that is, phrases with more than one lexical unit. Therefore, these also appear as part of the examples generated, as shown in Table 9: *A fuxida das/casas do concello* [The flight of the/council houses]; *A fuxida das/estacións de esquí* [The flight of the/ski stations]; *A fuxida da/residencia de estudiantes* [The flight from the/students' residence]. In other nouns such as *olor* [smell], there are also examples such as *cerveza rubia* [lager], *café con leche* [coffee with milk], *vino blanco* [white wine] or *licor café* [coffee liqueur]. However, dealing with these units from a typological point of view is not the goal at this work stage, since the research goal lies in the further description of the syntactic–semantic function of each lexical unit (simple or complex).

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

Table 2. Examples of cosine similarities between words

Word 1	Word 2	Cosine Similarity	Explanation
cat	cat	1.000	Perfect similarity, identical words
cat	dog	0.798	Substantial similarity between pets
cat	orange	0.263	Lower similarity, conceptually dissimilar
apple	orange	0.764	Higher similarity, both are fruits

Table 2 shows some examples of the cosine similarity between pairs of words. Note that we can only speculate about why the cosine similarity of two words is higher or lower. The properties assimilated by the embeddings model are not defined *a priori* but computed by the algorithm taking into account word contexts.

The similarity percentage used in our application is not a fixed threshold. To evaluate the model's performance, a series of human qualitative analyses were conducted using the same input data and its results for different thresholds between -1 (opposite) and 1 (identical)—see the *Combinatoria* tool. However, the results were inconclusive in determining a default sensitivity for all languages. Therefore, it is left to the user to decide the optimal threshold for each combination. Users are free to experiment with this functionality.

3.2 Local methodology for Galician and Portuguese

The first step is to identify the valency patterns for Galician and Portuguese nouns. On the basis of corpus analysis,[x] it is established which semantic categories together with their lexical representatives can fill the different valency slots. For this purpose, an ontology is applied (Domínguez Vázquez *et al.* 2021).

After some initial tests, we determined that the semantic classes observed for the Spanish nouns were coincident with the semantic categories in the equivalent Portuguese and Galician nouns. The prototypical semantic classes for the Spanish noun *mudanza* [move], such as semantic class {*lugar, construcción*}[place, building], were also prototypical in the case of Galician (*mudanza*) and Portuguese (*mudança*). Hence, this is to be expected given the proximity between languages and the application of the concept of a network of scenes (see Section 2.1).

A priori, nothing would hinder applying the methodological approach followed in *Xera*, but we decided to change it due to the following reasons:

- Repeating for each language the whole process of matching the valency argument, semantic classes and their lexical entries, as well as debugging inconsistencies, is time-consuming and not sustainable in multilingual projects.
- The outcomes of our interlanguage matching, as mentioned before, were solid.

From this, we considered another method to obtain the vocabulary required by the onomasiological tool *XeraWord*. Given that the matches between Spanish–Galician and Spanish–Portuguese were reliable, we decided to develop a tool for automatic translation of the lexical flow, *TraduWord* (Section 4.4). There is currently nothing similar. That is, in WordNet, we can obtain through the Interlingual Index (ILI) the translation of a specific word, but we cannot extract the translation of all the vocabulary linked to a semantic category, for example, all the vocabulary in

[x]For Galician: CORGA = Centro Ramón Piñeiro para a investigación en humanidades: Corus de Referencia do Galego Actual (CORGA) [3.2] http://corpus.cirp.gal/corga/. For Portuguese: Sketch Engine: Portuguese Web 2020 (ptTenTen20) https://www.sketchengine.eu/.

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

different languages linked to the semantic class {*lugar, construcción*} [place, building]. *TraduWord*, therefore, translates the vocabulary associated with a semantic category and its output data is therefore semantically prototyped. This makes possible the integration of these data into the new onomasiological tool of automatic generation for both target languages.

At the same time, working with WordNet had both a disadvantage and an advantage:

- We understand as an advantage that the vocabulary lists, which are already ontologically annotated and hierarchised into semantic classes, are linked to WordNet-ILIs. This would subsequently enable the translation from these languages into other languages using *TraduWord* and thus further develop a more comprehensive multilingual onomasiological resource.
- As a disadvantage, we assessed the limited size of GalNet and PULO (Section 4.1), which could affect the data variability and the quality of the generator. Hence, some additional experiments were performed using translation memories and machine translation tools, as will be described in the following sections. An analysis of the quality of the translations and the establishment of quality parameters was also necessary (Section 4.5).

Given the new onomasiological structure of the tool and the variation in the methodological process, for *XeraWord* a specific database is designed, a Lemmatiser and an Inflector are implemented, as well as application programming interfaces (APIs) and programming rules for the generation of both languages.

In summary, *XeraWord* is therefore different from the *Xera* tool. This is already reflected in the onomasiological structure of the first tool mentioned above.

4. Resources and tools for Portuguese and Galician

This section explains the methodology implemented to include the Galician and Portuguese languages in the original *Xera* tool, as well as the tools developed specifically for the semi-automatic translation of the required lexical resources. The section concludes with an analysis of the obtained results.

4.1 GalNet and PULO

Considering that *Xera* uses WordNet for concept mapping, the first decision was to resort to the Galician and Portuguese WordNets available in the Multilingual Central Repository[y] (MCR), as both resources are aligned at the sense level with the Princeton WordNet using the ILIs:

- GalNet[z] (Gómez Guinovart, 2011) has been developed for some years at the University of Vigo. While its bootstrap method was automatic and new term candidates are obtained by automatic techniques, their inclusion in the lexical database is validated manually. It comprises 42,724 synsets, divided into 31,759 nouns, 6,436 adjectives, 3,448 verbs and 1,081 adverbs. These synsets account for 67,978 different variants.
- PULO[aa] (Simões and Gómez Guinovart, 2014) has been under development for five years at the Polytechnic Institute of Cávado and Ave and University of Minho. Given the lack of human resources, the database was created using automatic methods, and no manual revision was performed on the data. At the moment it comprises 17,942 synsets, 10,047 of which are nouns, 3,581 adjectives, 3,786 verbs and 528 adverbs. These synsets account for 32,603 different variants.

[y] Available at http://adimen.si.ehu.es/web/MCR.
[z] Available at http://sli.uvigo.gal/galnet/
[aa] Available at http://www.wordnet.pt/

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

```
[
  {"form": "nena",  "lemma": "neno", "pos": "N", "gen": "F", "num": "S"},
  {"form": "nenas", "lemma": "neno", "pos": "N", "gen": "F", "num": "P"},
  {"form": "neno",  "lemma": "neno", "pos": "N", "gen": "M", "num": "S"},
  {"form": "nenos", "lemma": "neno", "pos": "N", "gen": "M", "num": "P"}
]
```

Figure 4. Sample JSON output for the inflexion tool.

It is important to note that these resources are quite smaller than the Princeton WordNet, which contains more than 100,000 synsets. This difference results in a relevant amount of senses that are not accounted for in these two resources. Therefore, a translation tool was also used. For this process, we chose to use the MyMemory[ab] service. Its choice was based on different factors: (i) it is free and has an easy-to-use API, and (ii) it is based both on human-created translation memories and, when those are unavailable, in machine translation algorithms.

4.2 Lemmatiser and inflector

We developed an *ad hoc* web-based inflexion API[ac] based on FreeLing's[ad] (Padró, 2012) dictionaries. The system uses a RESTful approach. The endpoint receives two mandatory fields: the language (`lang` field) and the lemma being inflected (`lemma` field).

Consider the following Hypertext Transfer Protocol (HTTP) request, using the command line tool `curl`, to get the inflexions for the Galician word *neno* [child]:[ae]

```
curl "http://ilg.usc.gal/flexionador/api?lemma=neno&lang=gl"
```

The web service answers with a JavaScript Object Notation (JSON) document with a list of forms. Each form includes its part of speech (in this case, noun) and its morphological properties (gender and number). Figure 4 shows the JSON document returned by the presented HTTP request.

4.3 Lexical packages

During the development of the original *Xera* tool, it was necessary to represent data regarding each argument structure and to establish the descriptive levels required for the proper functioning of the generator. To facilitate the interchange of these data between the system and the linguists, a tabular format was defined, so that these data could be exported from the database into Excel files to be manually edited and, later, reimported into the database.

These spreadsheets are composed of two distinct zones: a header and a body (Figure 5 presents an example of such a document):

- The header includes a unique identifier for the package, and all the remaining necessary information about the pattern of the structures to be generated, as well as keywords, ontological information and a representation of use case examples.
- The main body of the spreadsheet is composed of the lexical units associated with an ontological class and their inflexions. For each lemma, there is information about the Synset it was retrieved from.

[ab] Available at http://www.mymemory.com/.
[ac] Available at http://ilg.usc.gal/flexionador/apidoc.html.
[ad] Available at http://nlp.lsi.upc.edu/freeling/node/1.
[ae] The web service available at this address is only able to answer to Portuguese and Galician requests.

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

Table 3. Data in the current pilot phase of *XeraWord*

	Macro syntactic–semantic		
	Lexical packages	Frames: mono-argumental	Lemas: lexical units
Portuguese	242	53	5713
Galician	293	52	6382

```
## EMPAQUETADO
## Núcleo: mudanza en singular
## Anotación semántica: animado humano condición humana habitante
## Estructura argumental: determinante + {adjetivo_o} + nucleo + {adjetivo_o} + de + determinante +  actante
## Actante: N1
## lang: es
## ejemplo: la {inesperada} mudanza {inesperada} de los residentes
## action: compartido
## query1: ontology=sumo&category=SocialRole#SocialRole
## query2: ontology=epinonyms&category=ili-30-09620078-n&subcategories=on
## query3:
## query4:
## query5:
http://portlex.usc.gal/develop/combina.php?lang=es&action=compartido&query1=ontology%3Dsumo%26cat
egory%3DSocialRole%23SocialRole&query2=ontology%3Depinonyms%26category%3Dili-30-09620078-
n%26subcategories%3Don&query3=&query4=&query5=
aldeano     aldeano     N  M  S  10753442-n
aldeanos    aldeano     N  M  P  10753442-n
aldeana     aldeano     N  F  S  10753442-n
aldeanas    aldeano     N  F  P  10753442-n
ciudadano   ciudadano   N  M  S  10719267-n
ciudadanos  ciudadano   N  M  P  10719267-n
ciudadana   ciudadano   N  F  S  10719267-n
ciudadanas  ciudadano   N  F  P  10719267-n
inquilino   inquilino   N  M  S  10700517-n
inquilinos  inquilino   N  M  P  10700517-n
inquilina   inquilino   N  F  S  10700517-n
inquilinas  inquilino   N  F  P  10700517-n
```

Figure 5. Part of one lexical package for the structure N1 of the Spanish noun *mudanza* [move].

The current pilot phase of *XeraWord* provides data for five nouns in both Galician and Portuguese at different levels (see Table 3). This table presents the total number of packages, frames and lexical units that are registered for these five nouns.

Using Galician as the example language, an analysis of the database lets us see that each macro syntactic–semantic frame can be performed on an average of 5.56 lexical packages, meaning a similar number of semantic classes. Table 3 shows that a significant number of lexical units are analysed from a relational and ontological point of view.

For illustration purposes, follows an example of a frame for the Galician term *fuxida* [escape]:

$$\left[\ determinant + \{adjective_o\} + head + \{adjective_o\} + \text{“}ata\text{”} + determinant + actant\ \right]$$

Table 4 presents the different description levels of the noun *olor* [smell], which has eight syntactic–semantic frames and eighty-eight developed lexical packages, yielding an average of eleven lexical packages and semantic classes per mono-argumental pattern.

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

Table 4. Example for the description levels of a lemma

Description Levels		
Lemma level	olor – Sg.	
Quantitative internal level	Lexical packages	88
	Number of frames	8
Syntactic–semantic level	Frame structure example	*determinant* + {*adjective*$_o$} + *head*+
	(argument pattern)	{*adjective_o*} + *"de"* + *determinant* + *actant*
Semantic level	Semantic role	1: that which emanates or has an odour
	Ontological features	[animate], [building]
Morphosyntactic level	Syntactic function	subject
	Surface realisation	*"de"* + *determinant* + *noun*

Consequently, for the description of a structure such as the one shown in Table 4,

$$\left[\ determinant + \{adjective_o\} + head + \{adjective_o\} + \text{“}de\text{”} + determinant + actant\ \right]$$

for the semantic role 1, we obtain a broad set of lexical packages that are described from a syntactic point of view (as shown in Figure 6) but also from a bi-level semantic description: relational and ontological. Figure 6 shows the ontological categorisation for role 1 in the described argument pattern.

4.4 Package translation tool

TraduWord is the tool responsible for the automatic handling of all translations from Spanish to Galician and Portuguese. Concretely, it deals with the automatic translation of complete lexical packages for the arguments of a specific noun. *TraduWord* consists of the following steps:

(1) Translations are computed from the original Spanish lexical package using two different approaches: (i) Given the InterLanguage Indexes present in the lexical packages, which identify unambiguously a concept, the Portuguese WordNet's API is queried.[af] Note that although the service is available from the Portuguese PULO project, it returns results for every language available in the aforementioned MCR. The obtained results are used to create the list of variants for each synset present in the lexical package, for every available language. For the specific task being described, only the Portuguese and Galician words are considered; (ii) if no available results are coming from WordNet, the MyMemory service is used as a backup. For each Spanish term in the original lexical package, a request is performed to the MyMemory API to gather its translations in Portuguese and Galician.

(2) The obtained words are inflected using the API presented in Section 4.2. These data are stored in the database to guarantee that the generator will be faster, not needing to inflect words at run time.

(3) The last step consists of building a spreadsheet that can store all the computed data, that is, the number of terms translated, inflexion, and source of the translation. Figure 7 shows the beginning of the spreadsheet for the lexical package referring to the Portuguese lemma *presença* [presence].

[af] Available at http://wordnet.pt/api

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

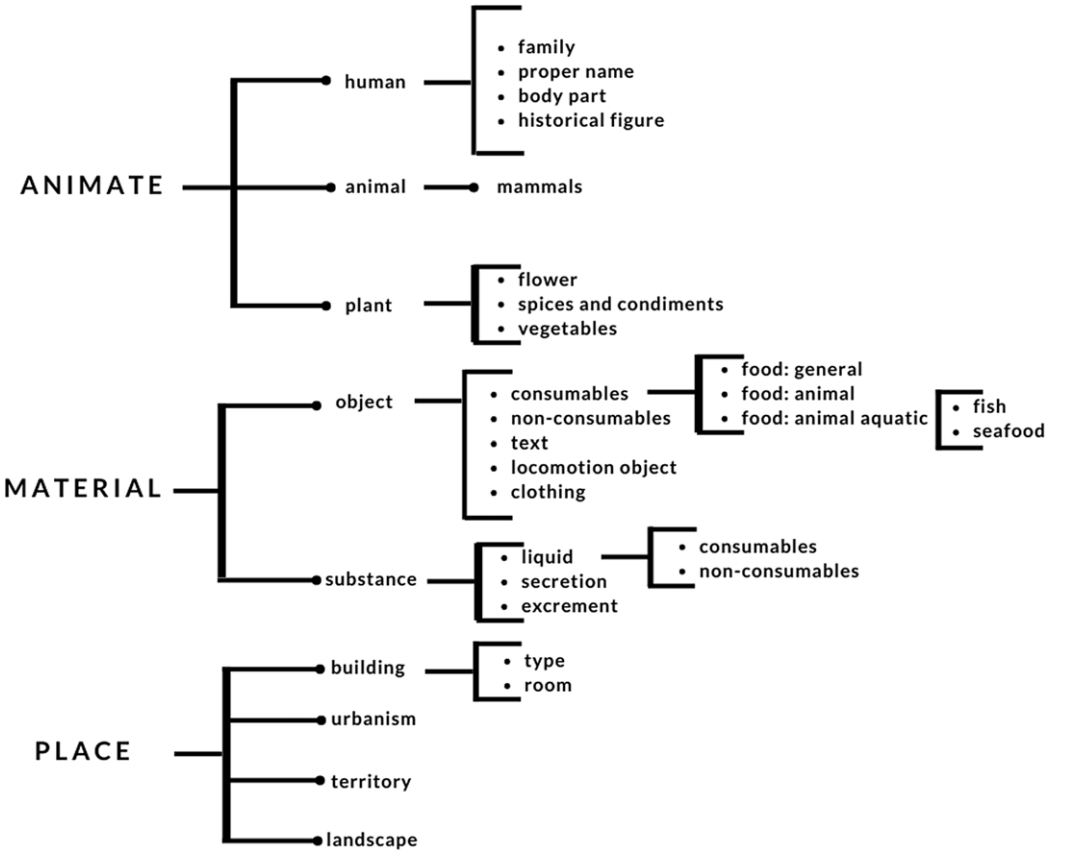

Figure 6. Ontological features with the semantic role AGENT for the Galician noun *olor* [smell] using the structure [*determinant* + {*adjective*$_o$} + *head* + {*adjective*$_o$} + *"de"* + *determinant*]

B	C	D	E	F	G	H
lemma	pos	gen	num	ili	source	lema_spa
marinha	N	F	S	08191701-n	mymemmory	armada
marinha	N	F	P	08191701-n	mymemmory	armada
exército	N	M	S	08199025-n	mymemmory	ejército
exército	N	M	P	08199025-n	mymemmory	ejército

Figure 7. Example of the spreadsheet generated for the lexical package referring to the Portuguese noun *presença* [presence].

These spreadsheets are manually reviewed to ensure the quality of automatic translations. It is also possible that the tool could not find the term in the target language and it must therefore be added manually. We are working closely with the authors of PULO and GalNet to guarantee that all the manual work performed for *TraduWord* can be used to enlarge these projects' lexical coverage.

The *TraduWord* framework can also be applied to other languages. The translation methodology for isolated words would not vary when translating into other languages, as long as these are available in the MCR. Thus, starting from a lexical package available in one of the languages of the *Xera* and *XeraWord* projects, the WordNet API could be used to obtain translations, if necessary,

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

Table 5. Percentage of terms translated with *TraduWord*

	WordNets	MyMemory	Combined
Portuguese	17.65%	81.29%	98.94%
Galician	43.26%	55.89%	99.16%

Table 6. Typology of problems found in the translations obtained from MyMemory

Type of problem	Spanish word	Translation into Portuguese
Words translated both into Portuguese and English	*masacre*	*massacre/ slaughter*
The translation does not match the meaning of the original word	*rosado* (wine)	*rosa/ cor de rosa* [colour]
Spelling problems	*gasóleo*	*6asole0*
Unnecessary specifications	*cazón*	*cação (peixe)* [extra fish note]
Changes in word class	*panameño* [an adjective]	*Panamá* [a noun]
Incorrect translation	*minino* [cat]	*meu filho* [my son]
Changes in the adjective's gender	*venezolano*	*venezuelana*
Scientific name for some animals	*pollo*	*gallus gallus/ frango*
Noise in the translations	*policía secreta*	*cone crazy/ polícia secreta*

complemented with data from MyMemory. Difficulties in the translation process into other languages may depend particularly on two factors: (i) the low number of variants for a specific synset in a given language or the absence of results from WordNet and (ii) translation problems involving linguistic aspects, such as polysemy, homography and inflexional problems. Specific problems may be expected depending on the language pair being translated. For example, in the case of the Spanish–German pair, one might expect the translation tool to present more difficulties in the case of multi-word units or compounds. Nevertheless, these problems do not depend only on the language itself but also on the results provided by WordNet and MyMemory. The following section discusses the problems that were observed when using the tool for the Portuguese and Galician translations. This analysis can be used as a starting indicator when working with other languages.

4.5 TraduWord: Analysing the data

The combination of WordNet and the MyMemory API allowed *TraduWord* to produce translations for most of the data. The addition of MyMemory was crucial for the automation of the translation process. This is especially the case for Portuguese, where the WordNet database is considerably underdeveloped in comparison with the other languages available in the MCR database. Despite the bigger size of the Galician WordNet, it covers less than half of the synsets from Princeton WordNet. Therefore, in both cases, MyMemory was responsible for translating over 50 per cent of the terms (Table 5).

The quality of the translations provided from these two data sources, however, is not the same and the data need to be analysed in detail. When manually reviewing the automatic translations obtained from MyMemory, some corrections or modifications had to be carried out. Table 6 highlights some of the problems found during the analysis of these translations.

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

Table 7. Distribution of translations according to their source for the Portuguese noun *cheiro* [smell]

	MyMemory		WordNet		Manually translated	
	Count	Perc.	Count	Perc.	Count	Perc.
translations	591	68.32%	237	27.40%	37	4.28%

As observed in Table 6, the translation tool MyMemory sometimes provides two translations for the same word. In these cases, only one of the provided translations is adequate regarding the resource's objectives. The inadequate one may be, for example, an English word or a scientific term. This implies that the automatic translation tool, despite providing correct translations, also produces a certain amount of noise that must be eliminated manually. However, we understand that the usage of MyMemory allows us to increase the number of translated items and allows a quicker bootstrap process than performing the translation manually.

Furthermore, there are several problems that are specifically related to polysemous words. Since only isolated words (with no context) are translated, sometimes the translation does not match the desired meaning of the word. This is not the case with the data obtained from WordNet, as ILIs are used to retrieve terms from an aligned synset, but with MyMemory data, this type of case is unavoidable. We can illustrate this problem with the example of the word *rosado*, which we can find in the lexical package corresponding to drinkable liquids in Spanish. In this context, it refers to a type of wine, but MyMemory translated it into Portuguese as a colour. This is, of course, a problem related to the usage of collocations that machine translation is not able to deal with correctly. A solution could be the usage of a tagger to identify these kinds of constructs or even generic named entities. In other cases, the translator only provides an incorrect translation: for example, in the case of the Spanish word *minino* (a colloquial term for 'cat') is automatically translated into Portuguese as *meu filho* [my son] as if the original sentence was *mi niño*. When this happens, these errors are eliminated and the translations are carried out manually.

Finally, in other cases, some modifications had to be introduced in the translation results. For example, sometimes the translation tool returns a term with unnecessary specifications in the target language, as in the case of the Spanish word *cazón*, which is translated into Portuguese as *cação (peixe)* [sandbar shark (fish)], with a parenthesised expression indicating that it is a type of fish. Some other times, the tool provides a correct translation but with the wrong gender or number (such as *venezolano* [Venezuelan] that, when translated from Spanish to Portuguese, changes its gender, returning *venezuelana*).

To be able to provide some quantitative data about these corrections, Table 7 depicts an example regarding the lexical package for the noun *cheiro* [smell].

As shown in Table 7, from a total of 865 terms in Spanish, 27.40 per cent of the translations into Portuguese have been obtained using WordNet data. None of these translations had to be modified. On the other hand, we can also see that most of the translations (68.32 per cent) were obtained from the MyMemory service. However, in this case, corrections had to be made in 19 (3.21 per cent) of the translations. In addition, 4.28 per cent of the terms were translated manually: this means that the original translations from MyMemory were not correct or did not match the meaning of the original word. This leaves a percentage of manual intervention higher than 6 per cent for this lexical package, but still acceptable when compared with the task of translating every word from scratch.

However, these data do not reflect a complete view of the manual intervention required to create complete lexical packages. In addition to the above-mentioned modifications and corrections, in many cases, some translated terms were removed manually. This was been done in the scenarios described below:

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

Table 8. Manual intervention in the translations according to their source for the Portuguese noun *cheiro* [smell]

	MyMemory		WordNet		MyMemory & WordNet	
	Count	Perc.	Count	Perc.	Count	Perc.
Correct translations	1761	38.98%	632	68.77%	2393	44.95%
Modifications	61	1.38%	0	—	61	1.14%
Deleted translations	2583	58.64%	287	31.23%	2870	53.91%
Total	4405		919		5324	

(1) When the translation tool offered several translations, some of them were incorrect. In this case, incorrect ones were eliminated.

(2) When the translation tool offered more than one translation for the same word (usually a polysemous one) and only one of the translations was the one desired for the specific context. This situation does not mean that the translation is incorrect, so it cannot be used as an indicator of the quality of the translation service.

(3) When a word is not a lexical representative to be included in a given lexical package in Portuguese but was in Spanish. It may be that, in a particular structure, the appearance of a given lexical unit is not frequent in Portuguese but is frequent in Spanish. This, again, is not an indicator of the quality of the translation service but is a normal process of linguistic revision that takes into account the peculiarities of each language.

Continuing with the example of the Portuguese noun *cheiro* [smell], we can obtain quantitative data to exemplify the level of manual intervention that has had to be carried out, considering also the terms that have been eliminated (see Table 8[ag]).

These data point to a lower accuracy of MyMemory when used as a translating tool since only 40 per cent of the translations are accounted for as correct. However, one observation must be made: MyMemory offers several translation alternatives for polysemous words. This means that unwanted alternatives for a specific context are eliminated, but the translations were not necessarily incorrect. In addition, MyMemory gives good coverage, as it allows us to automatically translate a larger number of terms.

WordNet, on the other hand, provides higher quality translations, even if it does not translate such a high number of terms. The elimination of translations provided by WordNet always responds to scenarios described in 2) and 3). That is to say, these are cases in which several alternatives were offered or where there were lexical units that were not representative in Portuguese for a particular lexical package.

Given the above, it can be concluded that these data allow us to measure the amount of manual work that had to be performed in the revision of the obtained translations. Although the effectiveness of translation tools is high, sometimes the special characteristics of the language make this manual work unavoidable. If, as in this case, terms have to be translated without being embedded in a given context (in a particular sentence, e.g.), we cannot prevent translation tools from considering all the meanings of a word and offering several alternatives.

[ag]The variation from the number of translated items we had in Table 7 has to do with the fact that, in this case, the inflexions have also been taken into account. In the previous case, the accounting was done based on the ILI, so that each lexical unit was considered only once, not taking into account their inflexions. Thus, for example, a word like the Spanish *niño* [boy] was only counted once, as a lemma. In this table, this same unit would be counted four times, one for each one of its inflexions: *niño, niña, niños, niñas*.

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

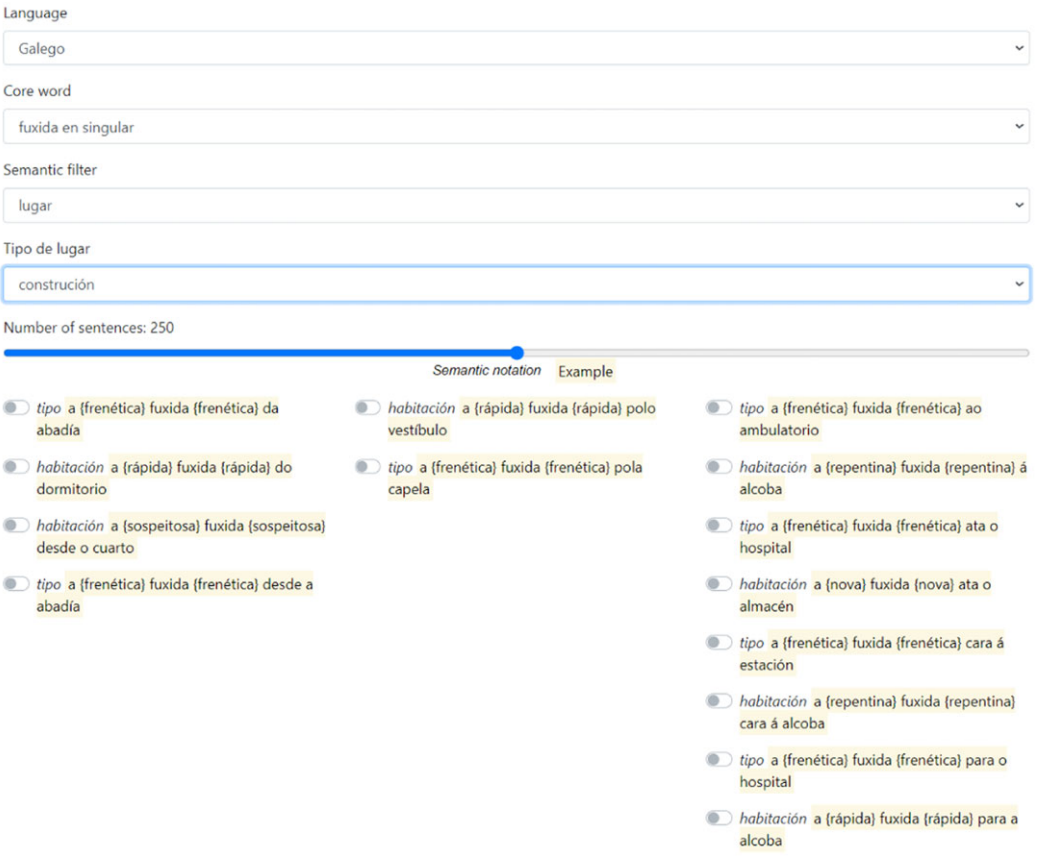

Figure 8. Query interface of *XeraWord*.

5. *XeraWord*: Implementation and visualisation

An API and a website were deployed to the website of the Instituto da Lingua Galega (ILG) to allow the use of these resources.

The interface for Galician and Portuguese was built from scratch to implement a semantic approach in the data visualisation that facilitates the generation process. This approach hides the complex metalinguistic information displayed on the original *Xera* application, such as argument structures or displaying the project metalinguistic ontological classification.

Before generating nominal examples, the user is free to choose one or several of the lexical packages presenting each structure in a different column. Finally, an example is presented for each surface realisation to help the user understand what will be generated (Figure 8).

Thus, to access the lexical data, *XeraWord* follows an onomasiological approach for the visualisation of this lexical data. The user is asked first to select the language and noun. Once this is done, the onomasiological information is presented to the user, who can select among the different argument realisations of the chosen ontological category. For instance, {*humano*} [human], {*lugar*} [place] or {*evento*} [event]. As indicated, the resource does not aim to provide data on collocations, but on valency complements and patterns. However, once the semantic classes have been displayed, the user can observe in the sample information on frequent non-valency combinations, that is, frequent collocations labelled with the symbol `{}`. In Figure 8 they are exemplified as follows: *Habitación - a {rápida} fuxida {rápida} do dormitorio* [Room - the {quick} escape {quick} from the bedroom]. In addition, some distributional information is given, as some adjectives can appear in Galician only before the noun, only after it, or in both positions. By clicking on the

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

bottom "Xerar Exemplos" [generate examples] the resource will then create phrases based on the selected criteria. Figure 8 shows this interface. The results can then be visualised on the website itself, or downloaded in either a comma-separated value file (CSV) or a JSON file.

The onomasiological approach implemented in *Xera* enables the visualisation of the formal representation of a semantic class. As shown in Figure 8, the interface presents the construction {*lugar, construcción*} [place, building] and shows that, in Galician, the place someone escapes from may be expressed with different prepositions depending on the action that is intended to be emphasised: *de* to highlight the place being escaped from, *por* referring to the itinerary or path taken and *para* and *a* are used for direction or destination. If a user selects, for example, the semantic feature *{tipo}* [type] for expressing the starting point of the escape with the prepositional realisation *de* in Galician and Portuguese (see Figure 8), *XeraWord* provides a list of automatically generated examples such as the ones shown in Table 9.[ah]

Despite the simplicity of the results, it is worth mentioning again that all the combinations generated respect the user's choice and several filters that may not be explicit to the user in the immediate visualisation of the data. This means that all the lexical candidates for the selected structure convey, together with the predicate (the place someone runs away from) and they are all examples of the ontological feature {*tipo, construcción*} [type, construction]. Furthermore, the user was able to observe and choose the desired surface realisation to express this information.

The output of *XeraWord* is always correct due to two reasons:

- In the first research phase, it was checked that *XeraWord* generated the data correctly following formal parameters.
- In a second analysis stage, the semantic data were processed: since the lexical packages and also their automatic translations (see Section 4.5) are checked manually, no error has been observed by the output.

In summary, if a) the grammatical-formal structure is programmed, b) it is established that a specific valency slot is filled by a specific semantic category, for example [place construction], and c) the semantically annotated, translated and checked vocabulary is right linked to this slot, as is the case, there is no margin for error for mono-argumental patterns. In any case, in further development of *XeraWord*, word embeddings can be implemented. This has two positive effects:

- Word embeddings enable the establishment of the most frequent co-occurrences, that is, to selectively generate only the most frequent mono-argumental valency patterns. This can be interesting, for example, for language learning.
- If the number of surface arguments increases, that is, when generating bi-argumental structures, semantic incompatibility problems may arise, for example, the woman's testicle pain. To avoid the generation of such examples, Word2vec (Mikolov *et al.* 2013) or FastText models (Bojanowski *et al.* 2017) can be implemented. Word2vec was successfully implemented in the *Combinatoria* tool.

This is a novel approach in the field of valency resources. This tool serves as the starting point for the future development of semasiological and combinatory-oriented dictionaries. As mentioned by Bosque and Mairal Usón (2012) and Simone (2012), most lexicographic resources follow an alphabetical organisation and ignore a native speaker's knowledge of combinatory or word semantic compatibility. Therefore, we need new resources that encode this knowledge to allow language learners to access and understand how to combine words, guaranteeing semantic correctness.

[ah] Depending on the query filters selected by the user the examples are generated randomly and independently for each language. Both tables are placed in parallel for space reasons, although they do not reflect the aligned translation of the examples.

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

Table 9. Examples generated for Galician and Portuguese

Galician	Portuguese
tipo	tipo
a {frenética} fuxida {frenética} da abadía	a {silenciosa} fuga {silenciosa} da prefeitura
A fuxida dos currais	A fuga das dependências
A fuxida dos moteis	A fuga das clínicas
A fuxida da pista	A fuga da residência
A fuxida da residencia	A fuga do sanatório
A fuxida do dispensario	A fuga das abadias
A fuxida do santuario	A fuga dos cinemas
A fuxida dos consulados	A fuga das igrejas
A fuxida do conservatorio	A fuga da embaixada
A fuxida da taberna	A fuga da clínica
A fuxida dos casinos	A fuga das moradias
A fuxida dos manicomios	A fuga da capela
A fuxida das enfermerías	A fuga do hospital
A fuxida das embaixadas	A fuga das prefeituras
A fuxida da casa do concello	A fuga do albergue
A fuxida da residencia de estudantes	A fuga do hospital
A fuxida da mesquita	A fuga dos teatros
A fuxida do bordel	A fuga das embaixadas
A fuxida do cibercafé	A fuga das tabernas
A fuxida das casas do concello	A fuga da residência
A fuxida dos ambulatorios	A fuga dos sanatórios
A fuxida das estacións de esquí	A fuga das capelas
A fuxida da mesquita	A fuga das barracas
A fuxida do bordel	A fuga do consulado

6. Conclusions

The main purpose of *XeraWord* as well as the original *Xera* and *Combinatoria* is to be used in the classroom as a new tool for L2 teachers (second language teachers) and learners alike. As of right now, however, the list of nouns implemented in this prototype is not enough to be used as the base of new didactic tools for the classroom.

Nevertheless, it can already be used for very different applications. Language dictionaries, which usually include one or two examples, can now redirect the dictionary user to a set of dynamically generated examples. This is especially true given the current trend in digital dictionaries.

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

In this regard, the difference between the generators, lexicographic portals, and other dictionaries is that the latter offers static examples. Therefore, users can obtain more examples by consulting them on the portal itself (Das Online-Wortschatz-Informationssystem Deutsch des Leibniz-Instituts für Deutsche Sprache (IDS)[ai] or Oxford Learners Dictionaries[aj]) or through a link to examples on the web (Merriam Webster[ak]). Still, they cannot select specific examples to consult. Searching for examples using a generation tool allows the users to filter the example type of example according to its surface realisation and ontological features. This allows the users to understand whether certain combinations are possible (or not) so that they can be used in production or reception situations and also to discover (or confirm) how words are used and whether or not a given word can be used in a particular environment.

Unlike the original resources, *Xera* and *Combinatoria*, *XeraWord* provides the user with an onomasiological organisation of the data that makes it easier to understand the different surface realisations of a noun in the selected language.

The main contributions of this project are:

(1) The methodology resulting from the combination of argument patterns analysis and lexical prototyping with WordNet and *TraduWord* as sources for obtaining lexical data is a successful methodology that can quickly adapt the tool to new languages. This is especially valuable for languages with few digital resources and/or few funding opportunities, as the method and the data are reusable.

(2) Given that the prepared lexical packages are similar across different languages and that the concepts are correlated between languages by their ILI from Princeton WordNet, this tool can also be used to generate multilingual sentences, and also as a parallel corpora source, or even as a tool for teaching and learning simple translation mechanics. We have already used the data to propose online exercises (see Section 2.2).[al]

(3) Increasing the number of units will allow the implementation of the tool in educational settings, such as language teaching and learning. The goal of *XeraWord* is to become a new tool for L2 teachers and learners alike.

The main takeaways from this investigation are (i) the development of resources to dump onomasiological data and (ii) the successful blending of combinatorial semantics, lexicography and natural language processing techniques. *XeraWord* is a step forward in developing what we have called interactive automatic dictionaries. These dictionaries combine different information systems that offer the final user a richer interaction with the resource. Therefore, this prototype represents a kind of dictionary of the future.

> **Dictionaries of the future** are lexical or linguistic information systems in which existing lexicographic data are conflated, multilingualism and linguistic variety are entrenched, and which provide people, when they are confronted with gaps in their knowledge, with an answer as well as support in the writing and formulation processes of texts (Vigoni-Theses, 2018).

Beyond developing new dictionaries, some data supporting the generator have been incorporated to enrich PULO and GalNet (see Section 4.1). Since these data are categorised syntactically and relational ontologically, it is the starting point for developing an automatic, multilingual and semantic-ontological corpora annotator. The pilot experiments are very promising (Arias Arias, 2022).

[ai] Available at https://www.owid.de/.

[aj] Available at https://www.oxfordlearnersdictionaries.com/.

[ak] Available at https://www.merriam-webster.com/.

[al] In a webinar with teachers, we have checked the suitability of such exercises regarding the language level of the target users. http://portlex.usc.gal/webinario/

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

Acknowledgement. We thank you Ana Salgado for the comments on this article. This contribution has been developed within the framework of the research '*Ferramentas TraduWord e XeraWord: tradución de caudal léxico e xeración automática da linguaxe natural en galego e portugués*' as collaboration project between the University of Santiago de Compostela, Institut for Galician Language (ILG), University of Vigo and the Polytechnic Institute of Cávado and Ave (IPCA) and financed by the University of Santiago de Compostela in the programme 'Convocatoria de proxectos colaborativos para institutos de investigación da USC. 2020', and partly funded by Portuguese national funds (PIDDAC), through the*Fundação para a Ciência e Tecnologia* (FCT) and FCT/MCTES under the scope of the project UIDB/05549/2020. Some progress lies in the framework of the project PID2022-137170OB-I00 (Spanish Ministry of Science and Innovation, FEDER - A way to make Europe - and the Spanish State Research Agency).

References

Álvez J., **Atserias J.**, **Carrera J.**, **Climent S.**, **Laparra E.**, **Oliver A. and Rigau G.** (2008). Complete and consistent annotation of WordNet using the top concept ontology. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco: European Language Resources Association (ELRA), pp. 1529–1534.

Arias Arias I. (2022). *Anotação semântica (semi)automática de corpora: A frase nominal em alemão*, Master's thesis. Universidade do Minho.

Atkins B. T. S. and Rundell M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press (OUP).

Bardanca Outeiriño D. (2020). *Automatic generation of dictionaries*, Master's thesis. Universidade de Santiago de Compostela.

Bentivogli L., **Forner P.**, **Magnini B. and Pianta E.** (2004). Revising the wordnet domains hierarchy: Semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, USA. Association for Computational Linguistics, pp. 101–108.

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

Bosque I. (ed.) (2023). Redes: diccionario combinatorio del español contemporáneo. Ediciones SM: Madrid.

Bosque I. and Mairal Usón R. (2012). Definiciones mínimas. In González R., (ed), *Estudios de lingüística española. Homenaje a Manuel Seco*. Universidad de Alicante: Servicio de publicaciones, pp. 119–132.

Bovi C. D. and Navigli R. (2017). Multilingual semantic dictionaries for natural language processing: The case of BabelNet. In *Semantic Computing*. World Scientific, pp. 149–163.

Busse D. (2012). *Frame Semantik*. Berlin/Boston: de Gruyter.

Domínguez Vázquez M. J., **Bardanca Outeriño D. and Simões A.** (2021). Automatic lexicographic content creation: Automating multilingual resources development for lexicographers. In *Post Editing Lexicography. Proceedings of the eLex 2021 Conference*, pp. 269–287.

Domínguez Vázquez M. J. (2015). Wie wird Veränderung im Deutschen und im Spanischen ausgedrückt? Ein szenen- und valenzfundiertes konzeptuelles wortklassenübergreifendes Beschreibungsmodell. In Meliss M. and Pöll B., (eds), *Aktuelle Perspektiven Der Kontrastiven Sprachwissenschaft: Deutsch-Spanisch-Portugiesisch. Zwischen Tradition und Innovation*. Tübingen: Narr, pp. 97–125.

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), 71–80.

Domínguez Vázquez M. J., **Solla Portela M. A. and Valcárcel Riveiro C.** (2019). Resource interoperability: Exploiting lexicographic data to automatically generate dictionary examples. In Kosem I., Kuhn T. Z., Correia M., Ferreira J. P., Jansen M., Pereira I., Kallas J., Jakubíček M., Krek S., and Tiberius C. (eds), *Proceedings of the eLex 2019 conference: Electronic Lexicography in the 21st Century*, pp. 51–71.

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

Domínguez Vázquez M. J., **Valcárcel Riveiro C. and Bardanca Outeiriño D.** (2021). Ontología léxica. Santiago de compostela. *Available at:* http://portlex.usc.gal/ontologia/.

Dowty D. (1991). Thematic proto-roles and argument selection. *Language* **67**(3), 547–619.

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

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

Fillmore C. (1977). Scenes-and-frames semantics. In Zampolli A., (ed), *Linguistic Structures Processing*. Amsterdam/New York/Oxford: North-Holland, pp. 55–81.

Foley W. A. and Valin R. D. V. (1984). *Functional syntax and universal grammar*. Cambridge: Cambridge University Press.

Fuertes-Olivera P. A., **Niño Amo M. and Sastre Ruano Á.** (2018). Tecnología con fines lexicográficos: Su aplicación en los Diccionarios Valladolid-UVa. *Revista Internacional de Lenguas Extranjeras* **10**, 75–100.

Gao J. and Liu H. (2020). Valency dictionaries and Chinese vocabulary acquisition for foreign learners. *Lexikos* **30**(1), 1–32.

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

Gómez Guinovart X. (2011). GalNet: WordNet 3.0 do galego. *Linguamática* **3**(1), 61–67.
Gómez Guinovart X. and Solla Portela M. A. (2018). Building the Galician wordnet: methods and applications. *Language Resources and Evaluation* **52**(1), 317–339.
Hermanns F. and Holly W. (2007). Linguistische Hermeneutik. In *Theorie und Praxis des Verstehens und Interpretierens*. Berlin: de Gruyter.
Humblé P. (1998). The use of authentic, made-up and 'controlled' examples in foreign language dictionaries. In *Proceedings of the 8th EURALEX International Congress*, pp. 593–599.
Izquierdo Beviá R., Suárez Cueto A. and Rigau Claramunt G. (2007). Exploring the automatic selection of basic level concepts. In *Proceedings of the International Conference in Recent Advances in Natural Language Processing*, pp. 298–302.
Jacinto García E. (2015). *Forma y función del diccionario. Hacia una teoría general del ejemplo lexicográfico*. Jaén: Servicio de Publicaciones de la Universidad Jaén.
Kabashi B. (2018). A lexicon of Albanian for natural language processing. In Čibej J., Gorjanc V., Kosem I. and Krek S., (eds), *18th EURALEX International Congress: Lexicography in Global Contexts*, pp. 855–862.
Kipper K., Dang H. T. and Palmer M. (2000). Class-based construction of a verb lexicon. In *7th National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pp. 691–696.
Køhler Simonsen H. (2020). Augmented writing and lexicography: A symbiotic relationship? In *Proceedings of XIX EURALEX Congress*, vol. 1, pp. 509–514.
Kosem I., Koppel K., Kuhn T. Z., Michelfeit J. and Tiberius C. (2018). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography* **32**(2), 119–137.
Laufer B. (1992). Corpus-based versus lexicographer examples in comprehension and production of new words. In *Proceedings of the 5th EURALEX International Congress*, pp. 71–76.
Lettner K. (2019). *Zur Theorie Des Lexikographischen Beispiels: Die Beispielangaben in der ein- und zweisprachigen pädagogischen Lexikographie des Deutschen*. Berlin: de Gruyter.
Mel'čuk I. (1996). Lexical functions: a tool for the description of lexical relations in a lexicon. In Wanner L., (ed), *Lexical functions in lexicography and natural language processing*, volume 31 of Studies in Language Companion Series. Amsterdam: John Benjamins, pp. 37–102.
Meyers A., Reeves R., Macleod C., Szekely R., Zielinska V., Young B. and Grishman R. (2004). Annotating noun argument structure for NomBank. In *4th International Conference on Language Resources and Evaluation (LREC). European Language Resources Association (ELRA)*, pp. 803–806.
Mikolov T., Chen K., Corrado G. and Dean J. (2013). Efficient estimation of word representations in vector space, Published by arXiv, https://arxiv.org/abs/1301.3781v3
Miller G. A., Beckwith R., Fellbaum C., Gross D. G. and Miller K. (1990). Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography* **3**(4), 235–244.
Müller-Spitzer C., Nied Curcio M., Domínguez Vázquez M. J., Dias I. M. S. and Wolfer S. (2018). Rechercheprаxis bei der Verbesserung von Interferenzfehlern aus dem Italienischen, Portugiesischen und Spanischen: Eine explorative Beobachtungsstudie mit DaF-Lernenden. *Lexicographica* **34**(1), 157–182.
Moreno Jiménez L. G., Torres-Moreno J. M., S. Wedemann R. and SanJuan E. (2020). Generación automática de frases literarias. *Linguamática* **12**(1), 15–30.
Nallapati R., Zhou B., dos Santos C., Gulçehre C. and Xiang B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, Berlin: Association for Computational Linguistics, pp. 280–290.
Nied Curcio M. (2014). Die Benutzung von Smartphones im Fremdsprachenerwerb und -unterricht. In *Proceedings of the 16th EURALEX International Congress*, Bolzano, Italy: EURAC research, pp. 263–280.
Niles I. and Pease A. (2001). Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems*, New York, USA: Association for Computing Machinery, pp. 2–9.
Otter D. W., Medina J. R. and Kalita J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21.
Otter D. W., Medina J. R. and Kalita J. K. (2021). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems* **32**(2), 604–624.
Padró L. (2012). Analizadores multilingües en FreeLing. *Linguamática* **3**(2), 13–20.
Polenz P.v (2012). *Deutsche Satzsemantik: Grundbegriffe des Zwischen-den-Zeilen-Lesens*. Berlin: de Gruyter.
Prinsloo D., Heid U., Bothma T. and Faaß G. (2012). Devices for information presentation in electronic dictionaries. *Lexikos* **22**(1), 290–320.
Prinsloo D. J., Heid U., Bothma T. and Faaß G. (2011). Interactive, dynamic electronic dictionaries for text production. In Kosem I. and Kosem K., (eds), *Electronic lexicography in the 21st Century: New Applications for New Users (eLex2011)*, pp. 215–220.
Roemmele M. (2016). Writing stories with help from recurrent neural networks. In *Proceedings of The 13th AAAI Conference on Artificial Intelligence*, Phoenix, USA: AAAI Press, pp. 4311–4312.

https://doi.org/10.1017/nlp.2024.32 Published online by Cambridge University Press

Rundell M. (1998). Recent trends in English pedagogical lexicography. *International Journal of Lexicography* **11**(4), 315–342.
Ruppenhofer J., **Ellsworth M.**, **Schwarzer-Petruck M.**, **Johnson C. R. and Scheffczyk J.** (2006). *FrameNet II: Extended theory and practice*. Berkeley, CA: International Computer Science Institute, Technical report.
Simões A. and Gómez Guinovart X. (2014). Bootstrapping a portuguese wordnet from galician, spanish and english wordnets. *Advances in Speech and Language Technologies for Iberian Languages* **8854**, 239–248.
Simone R. (2012). *Diccionarios que todavía no existen. Lecture at the V Congreso internacional de lexicografía*. Madrid: Universidad Carlos III.
Smith C. (1991). *The Parameter of Aspect*. Dordrect: Kluwer Academic Publishers.
Spohr D. (2011). A multi-layer architecture for Pluri-Monofunctional dictionaries. In Fuertes Olivera, P.A, Bergenholtz, H. (eds.) *In e-Lexicography: The Internet, Digital Initiatives and Lexicography*. London: Bloomsbury Academic, pp. 103–120.
Vicente M., **Barros C.**, **Peregrino F. S.**, **Agulló F. and Lloret E.** (2015). La generación de lenguaje natural: análisis del estado actual. *Computación y Sistemas* **19**(4), 721–756.
Vigoni-Theses V. (2018). Dictionaries for the future, Available at https://www.emlex.phil.fau.eu/files/2019/03/Villa-Vigoni-Theses-2018-English.pdf.
Žabokrtský Z., **Zeman D. and Ševčíková M.** (2020). Sentence meaning representations across languages: What can we learn from existing frameworks? *Computational Linguistics* **46**(3), 605–665.
Zhu H., **Li Y. and Chiticariu L.** (2019). Towards universal semantic representation. In *1th International Workshop on Designing Meaning Representations*. Florence, Italy: Association for Computational Linguistics, pp. 177–181.
Ziem A. (2008). *Frames Und Sprachliches Wissen. Kognitive Aspekte der semantischen Kompetenz*. Berlin/New York: de Gruyter.

Cite this article: Domínguez Vázquez MJ, Simões A, Bardanca Outeiriño D, Caíña Hurtado M and Iglesias Allones JL (2025). Automatic generation of nominal phrases for Portuguese and Galician. *Natural Language Processing* **31**, 606–630. https://doi.org/10.1017/nlp.2024.32