Proposta recibida en xullo 2020 e aceptada para publicación en decembro 2020.

Aplicación de WordNet e de word embeddings no desenvolvemento de prototipos para a xeración automática da lingua

Application of WordNet and word embeddings in the development of prototypes for automatic language generation

María José Domínguez Vázquez
Universidade de Santiago de Compostela-ILG
majo.dominguez@usc.es

Resumo

Esta presentación de dous prototipos de xeración automática de lingua natural achega unha visión de conxunto da metodoloxía aplicada na descrición e procesamento dos datos lingüísticos, así como das técnicas e ferramentas xa existentes ou desenvolvidas co fin de garantir o funcionamento dos simuladores en alemán, español e francés.

Palabras chave

Gramática de valencias, Patróns argumentais, Procesamento da linguaxe natural (PLN) e Xeración da linguaxe natural (XLN), WordNet, word embeddings

Abstract

This presentation of two prototypes of automatic natural language generation provides an overview of the methodology applied in the description and processing of linguistic data, as well as of the techniques and tools already existing or developed in order to guarantee the functioning of the simulators in German, Spanish and French.

Keywords

Valency Grammar, Argument patterns, Natural Language Processing (NLP) and Natural Language Generation (NLG), WordNet, word embeddings

1. Introdución: estado do problema

Un importante hándicap na intelixencia artificial é converter a información lingüístico-semántica en información codificable e reproducible por non humanos, isto é, dotar as máquinas de coñecemento léxico (Navigli & Ponzeto, 2012) coa pretensión de que as computadoras interactúen coa súa contorna de modo humanizado utilizando as linguas naturais como código de intercambio.[1] Nesta liña, resulta especialmente relevante o deseño de léxicos computacionais — lexibles por máquinas e dotados de información semántica —, que faciliten a desambiguación dos diferentes significados (Agirre & Edmonds, 2006). As investigacións lexicográficas poden impulsar significativamente esta tarefa, se entendemos que "[b]y far the best existing semantic descriptions of language are dictionaries [...]" (Trap-Jensen, 2018, p.34). Nesta dobre aproximación — a lexicográfica e computacional — enmárcanse os prototipos de xeración automática de frases nominais en contexto para o español, francés e alemán, *Xera* e *Combinatoria* (*vid.* 2.2).[2]

Estes xeradores nacen ligados ao dicionario online multilingüe PORTLEX[3] , un recurso valencial, semicolaborativo, modular, multilingüe e *cross-lingual* sobre o potencial combinatorio da frase nominal en 5 linguas (Domínguez & Valcárcel, 2020). As principais dificultades no seu desenvolvemento non so atinxen á notable inversión de tempo na análise e compilación de todos e cada uns dos esquemas argumentais así como do seu potencial combinatorio (Domínguez & Valcárcel, 2020), senón tamén ás limitacións observadas no uso de córpora:[4]

- Na extracción dos datos tirados dos córpora non é posible aplicar ningún filtro de pre-

[1] A día de hoxe xa existen asistentes virtuais como Siri, Alexa ou Google Home, isto é, máquinas que "falan", pero tamén temos máquinas que aprenden e son adestradas.

[2] http://portlex.usc.gal/combinatoria. Os recursos son gratuitos e de libre acceso. Para máis información, *vid.* a información sobre as condicións de uso na anterior ligazón.

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

[4] Cómpre dicir que os córpora ofrecen paulatinamente vías máis precisas en prol da análise das propiedades distributivas e sintagmáticas do léxico. Serven de exemplo Sketch Engine (https://www.sketchengine.eu/), *COSMAS II* (https://cosmas2.ids-mannheim.de/cosmas2-web/), CORGA (http://corpus.cirp.es/corga/) ou TLFi (http://atilf.atilf.fr/).

DOI: 10.21814/lm.12.2.337

selección semántico-argumental nin un envorcado de datos atendendo as acepcións de significado.[5] Así, o lexema *x* no composto *x* + UMZUG alemán na súa acepción 'cambio de domicilio ou lugar de traballo' pode expresar un rol semántico AXENTE — "Aquel ou aquilo que realiza unha acción". Non obstante, a listaxe de frecuencia obtida mediante unha busca *Corpus Query Language* (CQL) en *Sketch Engine* recolle maioritariamente exemplos de compostos doutra acepción de significado, en concreto de "cabalgata" ou "desfile".[6] Xunto coa necesidade, pois, de cribar os compostos atendendo ao seu significado relacional, engádese a de atopar de xeito automático as súas posibles combinatorias cos outros actantes argumentais nas diferentes realizacións — combinatorias, ademais, comúns na lingua.

- Os resultados envorcados non cumpren os requisitos para a súa inclusión nun dicionario de valencias ao non exemplificar complementos argumentais. Un claro exemplo desta casuística son as realizacións adxectivais nas diferentes linguas contempladas no recurso. Unha análise semántica dos 70 primeiros adxectivos obtidos de *Sketch Engine* para o esquema argumental AUSENCIA + *Adxectivo*, permite concluír que soamente unha porcentaxe moi reducida serve para o propósito do noso recurso. De entre esta listaxe só un 4 % dos adxectivos representa un rol semántico e soamente un único exemplo pode ser adxudicado ao rol AXENTE (*a ausencia escolar*), sendo o mesmo, porén, ambiguo.

Xa que a frecuencia léxica no eixo sintagmático non está en necesaria correlación coa súa función específica ou argumental, atopar candidatos léxicos que cubran o actante valencial que se exemplifica supón a análise manual dun número significativo de coocorrencias, e isto, para as 5 linguas contempladas no dicionario. Constatadas estas dificultades, que ademais atrasaban enormemente o traballo, decidimos proceder á inversa: xerar directamente as estructuras argumentais aplicando previamente filtros semántico-combinatorios (*vid.* 2.2), no canto de buscalas nos córpora. Este é o punto de partida dos simuladores.

2. Os xeradores lingüísticos: metodoloxía e fases

2.1. Visión de conxunto

É a partir dos anos 90 cando se intensifica a investigación no campo da xeración automática da lingua natural. Nun principio os xeradores non priorizaban aspectos importantes como a integrabilidade, a portabilidade ou a eficiencia, nin lle dedicaban especial atención a aspectos de natureza máis lingüística, como a coherencia semántica ou textual, factores que, xunto coa fluidez e a variación nas posibles realizacións, son elementos centrais na avaliación dos xeradores de lingua natural (Hashimoto et al., 2019; Horacek & Zock, 2015; Jiménez et al., 2020; Vicente et al., 2015). Xunto con protocolos e estudos de avaliación da calidade dos resultados, xa se poden xerar de xeito automático textos e frases con diferentes características (Vicente et al., 2015; Nallapati et al., 2016; Sordoni et al., 2015) — incluso case sen input de partida (Roemmele, 2016) —, así como imaxes a partir de textos e viceversa (Otter et al., 2020). As aproximacións á xeración automática dende ou para a aplicación lexicográfica non é ampla.[7] Existen diferentes propostas para a xeración automática de dicionarios (Bardanca Outeiriño, 2020; Kabashi, 2018; Delli Bovi & Navigli, 2017), de artigos lexicográficos (Geyken et al., 2017) ou ben dalgunha das súas partes (os exemplos, en Kosem et al. (2019)). Estes recursos, porén, non perseguen os mesmos obxectivos que os xeradores que aquí presentamos.

2.2. Os xeradores *Xera* e *Combinatoria*

Os dous prototipos para a xeración de patróns argumentais de frases simples (*Xera*) e complexas (*Combinatoria*) en español, francés e alemán

[5] Recursos de diferente tipoloxía achegan descricións semántico-argumentais, en especial, para o inglés — *Verbnet* (http://verbs.colorado.edu/~mpalmer/projects/verbnet.html), Propbank (http://verbs.colorado.edu/~mpalmer/projects/ace.html), *VerbAtlas* (http://verbatlas.org/), *CPA* (http://www.pdev.org.uk/). Para outras linguas, como español e catalán, *vid.* AnCora (http://clic.ub.edu/corpus/es/ancora).

[6] Por cuestión de espazo amósanse soamente os 10 primeiros lemas e a súa frecuencia: *Festumzug* (18742), *Laternenumzug* (5568), *Faschingsumzug* (5191), *Karnevalsumzug* (4669), *Martinsumzug* (2838), *Rosenmontagsumzug* (2441), *Serverumzug* (1952), *Lampionumzug* (1638), *Fackelumzug* (1626) e *Fastnachtsumzug* (1467).

[7] Si que se aplican diferentes ferramentas de análise e técnicas, por exemplo, FreeLing (Padró, 2012), estudos para a extracción automática de diferentes tipos de datos (Gamallo & Pichel, 2007; Kilgarriff et al., 2008; Renau & Nazar, 2016), así como con softwares que permiten a compilación de dicionarios, como, por exemplo, TshwaneLex (https://tshwanedje.com/). Así mesmo, créanse aplicacións de dicionarios para dispositivos móbiles a partir de redes semánticas como WordNet, por exemplo o Dicionario GalNet (http://sli.uvigo.gal/digalnet/) de Gómez Guinovart en Google Play.

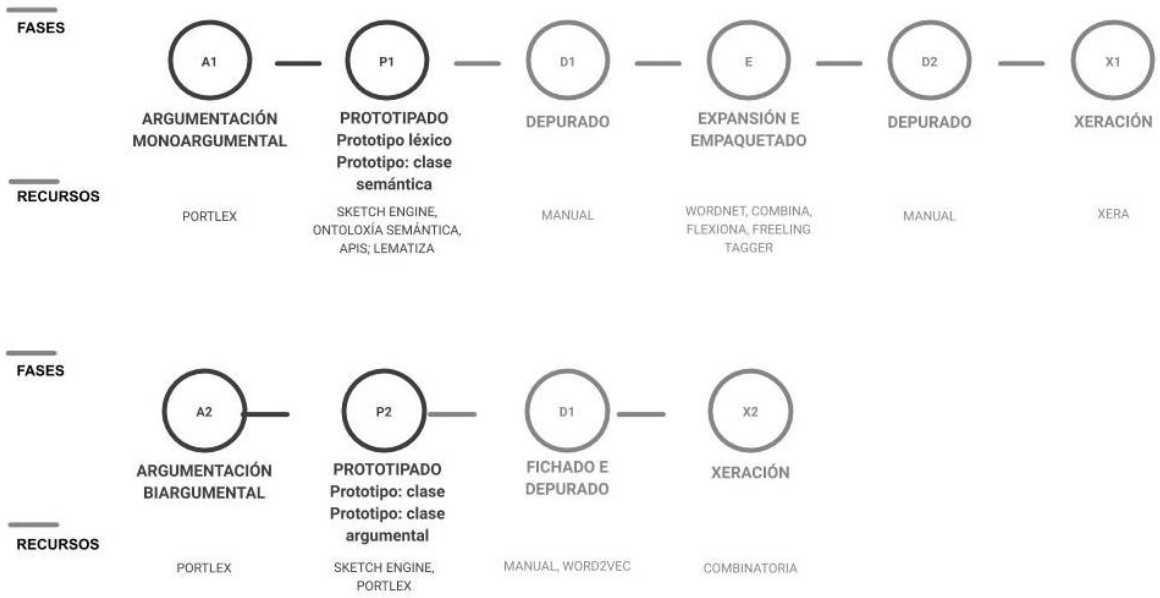

Figura 1: Metodoloxía e recursos de xeración automática argumental

están en funcionamento. Para o seu deseño precisamos unha metodoloxía que nos permitise describir e procesar os datos lingüísticos con información sintáctico-semántica combinatoria argumental, de tal xeito que estes se puideran programar e xerar automaticamente. Para tal fin, combinamos propostas teóricas e metodolóxicas recorrendo á gramática e semántica valencial, á teoría dos prototipos léxicos, ao PLN (recuperación e extracción) e XLN, así como a WordNet. Na segunda das ferramentas, aplicamos o método predictivo Word2vec (*vid.* 3). A Figura 1 recolle a metodoloxía xeral e as ferramentas que dan sustento aos xeradores no proceso de xeración da estrutura argumental; a Figura 2 resume o procedemento da contextualización, que xa se desenvolveu para a frase adxectival, pero está en curso no caso da oración.

O primeiro paso para a xeración automática da argumentación e o potencial combinatorio nominal consiste en establecer que actantes argumentais son específicos de cada substantivo[8] e que caudal léxico pode cubrir o eixo paradigmático dos devanditos actantes. Tomando como referencia a estrutura argumental do dicionario PORTLEX, analizamos os trazos ontolóxicos-categoriais (Engel, 2004) das (co)aparicións argumentais extraídas seguindo criterios de frecuencia de *Sketch Engine*. Atopámonos na fase de prototipado léxico: buscamos, por tanto, candidatos prototípicos para cubrir actantes funcionais concretos, isto é, exemplares léxicos, como, por exemplo, os lexemas *suor*, *tabaco* ou *pólvora* no rol CLASIFICATIVO de OLOR + *A*. Séguelle a segunda fase de prototipado, que consiste na determinación das clases semánticas prototípicas dun argumento concreto. Para o caso citado, por exemplo, [+Material] [+Substancia]: *suor*, *tabaco*, *pólvora* — [+Animado] [+Planta]: *flor*, *rosa* — [+Animado] [+Animal]: *porco*, *can* etc. A Figura 3 amosa un exemplo.

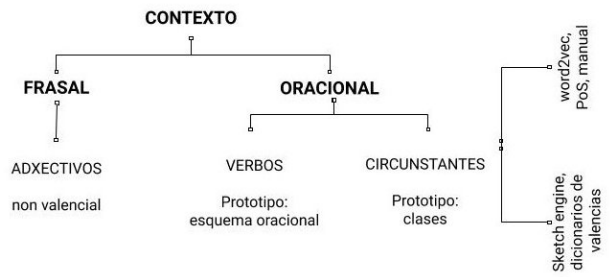

Figura 2: Metodoloxía xeral de contextualización

A descrición dos trazos categoriais e o prototipado léxico son conceptos esenciais non só dende un punto de vista puramente descritivo, senón tamén porque nos permiten acometer o seguinte estadio de traballo: a expansión léxica recorrendo a WordNet. Esta ampliación do número de candidatos léxicos atribuíbles a cada actante funcional conséguese mediante os trazos categoriais,

[8] A escolla dos 10 substantivos dos prototipos segue dous criterios centrais: (a) o seu estatus de portadores valenciais, isto é, a súa capacidade de abrir casillas funcionais e (b) a súa pertenza a diferentes campos semánticos, como Locación (*presenza*), Expresión (*pregunta*, *discusión* e *texto*), Afección (*morte*, *aumento* e *dor*) e Clasificación (*olor* e *sabor*). Deste xeito obtemos un amplo abanico de esquemas sintáctico-argumentais e combinatorias.

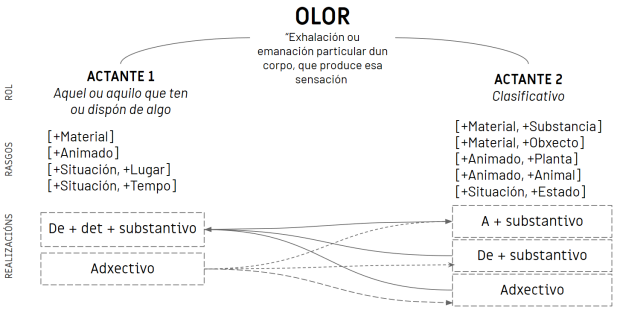

Figura 3: Argumentación e clases semánticas prototípicas

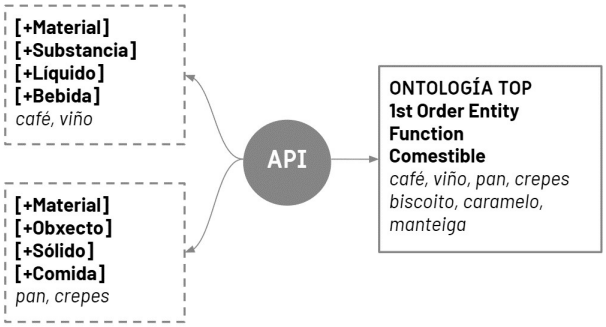

Figura 4: Expansión léxica usando WordNet

que fan de nexo de unión coas clases e atributos das categorías de WordNet (*vid.* Figura 4). Domínguez et al. (2019) indican que "the synsets of the wordnets following the EuroWordNet model of the Multilingual Central Repository (MCR) (...) are associated with semantic or cognitive features categorized in different ontologies".[9]

Aínda que esta organización cognitiva das ontoloxías de WordNet non coincide exactamente co inventario categorial da lexicografía valencial, tamén se constata que os trazos categoriais de tipo xeral si que conforman clases xerais nas ontoloxías do MCR. Por tanto, mediante este procedemento de expansión léxica[10], e posterior depuración e etiquetaxe seguindo unha ontoloxía sumativa de elaboración propia, acádase unha selección de caudal léxico no eixo paradigmático que comparte as características semánticas do prototipo léxico-semántico que tomamos como punto de partida, isto é, que conforma unha clase semántica concreta. Por tanto, a partir dos prototipos léxicos determinamos clases semánticas prototípicas. Unha vez obtido este caudal léxico, realízase a flexión e o empaquetado, no que temos en conta cuestións de etiquetaxe morfosintáctica e semántica. Estes repertorios léxicos atribuídos a cada argumento nominal son fundamentais para a xeración.

3. Ferramentas e recursos

No desenvolvemento dos xeradores operamos nas diferentes fases con (a) recursos existentes ou con ferramentas creadas *ad hoc* (*vid.* Figuras 1 e 5), entre as que diferenciamos (b) as que dan sustento aos xeradores, así como (c) os xeradores en si mesmos, que envorcan os datos en formato JSON e CSV.[11]

(a) **Recursos existentes:** Cabe subliñar aquí Sketch Engine, FreeLing e WordNet: i) **Sketch Engine** permítenos extraer, mediante consultas CQL, a frecuencia de cooaparicións en diferentes estruturas argumentais que tomamos do dicionario PORTLEX; ii) para o desenvolvemento do código de flexión partimos dos dicionarios do etiquetador **FreeLing**. Xa que na extracción léxica a partir de WordNet tamén se obteñen formas compostas — sendo estas unha posible realización argumental — recorremos tamén a este recurso para obter a división en lemas. iii) A rede semántica **WordNet** (Gómez Guinovart & Solla, 2018) posibilita a obtención do caudal léxico atendendo a clases ontolóxico-categoriais (*vid.* b).

(b) **Ferramentas para a análise lingüística desenvolvidas *ad hoc*:**[12]

- Coa finalidade de extraer datos léxicos das consultas que recorren ás relacións semánticas de WordNet e ás ontoloxías vinculadas aos synsets no modelo EuroWordNet desenvolvéronse tres **APIs**, unha para cada lingua obxecto de estudo.[13].

[9] Dado que ao inicio do traballo unicamente o español contaba cun wordnet vinculado ás ontoloxías mencionadas como parte do MCR (http://adimen.si.ehu.es/web/MCR), foi preciso crear bases de datos para o francés e o alemán. O procedemento explícase en Domínguez et al. (2019, p. 61–62): "This was done by extracting the alignment between lexical variants and identifying offsets of the meaning from the WordNet Libre du Français (WOLF) (Sagot & Fišer, 2008) and with data from the Extended Open Multilingual WordNet (Bond & Foster, 2013). Both have been made available on the GalNet interface after being converted to the EuroWordNet format of the MCR.". Dende o 2017 tamén está en desenvolvemento o Open German WordNet (https://github.com/hdaSprachtechnologie/odenet).

[10] Para máis información *vid.* a descripción da ferramenta *Combina* no apartado 3.

[11] Poden ser, por tanto, de aplicación como léxicos computacionais lexibles por máquinas con información semántica.

[12] Unha descrición detallada das ferramentas atópase en (Domínguez et al., 2019).

[13] http://portlex.usc.gal/develop/de/api/; http://portlex.usc.gal/develop/es/api/; http://portlex.usc.gal/develop/fr/api/

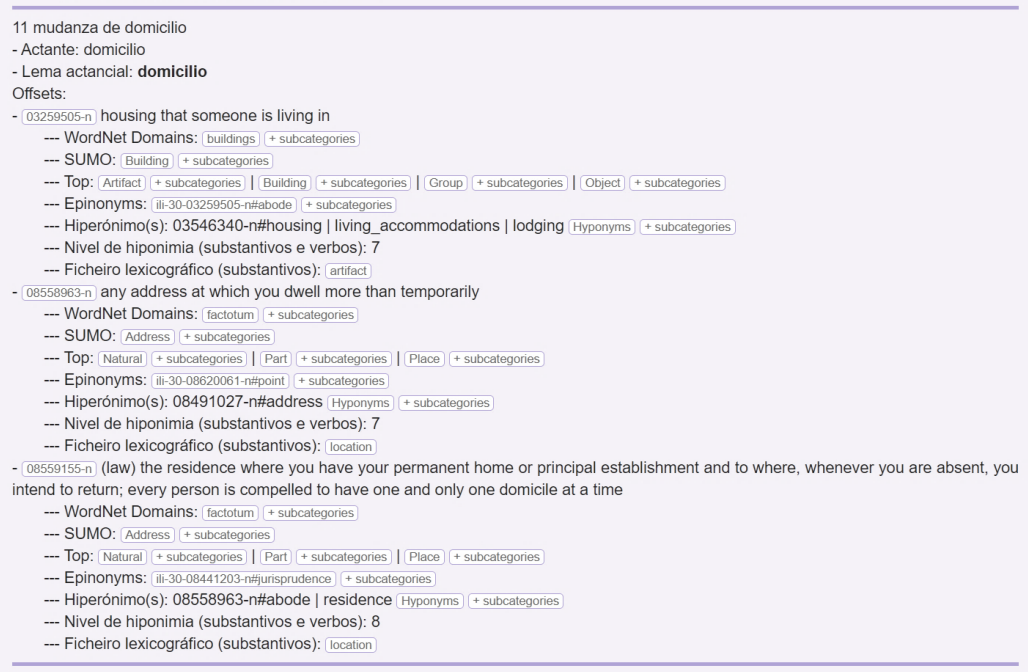

Figura 5: Imaxe de Lematiza

- A partir dos datos envorcados do corpus Sketch Engine, **Lematiza**[14] amosa os lemas coas súas variantes de significado (**synsets**) ligadas ás diferentes ontoloxías de WordNet (*vid.* Figura 5). Tamén é posible acceder a cada unha das categorías das ontoloxías de WordNet directamente premendo nos diferentes apartados.
- A ferramenta **Combina**[15] permite realizar consultas ás ontoloxías de WordNet (*vid.* 2.2) e, deste xeito, o envorcado semiautomático de datos compartidos ou combinados da *Top Concept Ontology*[16] (Álvez et al., 2008), dos *WordNet Domains*[17] (Bentivogli et al., 2004), da *Suggested Upper Merged Ontology*[18] (Niles & Pease, 2001), dos *Basic Level Concepts* (Izquierdo et al., 2007) e dos *Epinónimos* (Gómez Guinovart & Solla, 2018), ademais dos primitivos semánticos de Miller et al. (1990). Tras un procedemento de depuración e etiquetaxe, este caudal léxico expandido conforma os paquetes léxicos. Por tanto, se Lematiza amosa as variantes de significado (por exemplo para "domicilio" na Figura 5)[19], Combina presenta o repertorio léxico expandido no eixo paradigmático. Así, na busca de exemplares léxicos que, por exemplo, ocupen o actante locativo para un esquema argumental do tipo AUSENCIA + DE + *Locación* unha busca compartida nesta ferramenta[20] envorca un repertorio léxico en consonancia cos prototipos léxicos establecidos, por exemplo:
  - `[03259505-n` casa`]`,
  - `[04172107-n` chalet adosado`]`,
  - `[03259505-n` domicilio`]`,
  - `[03259505-n` piso`]`,
  - `[03259505-n` residencia`]` ou
  - `[04517408-n` segunda residencia`]`.
- A ferramenta **Flexiona** realiza a flexión dos lemas seleccionados.

(c) **Ferramentas de xeración: *Xera* e *Combinatoria*.**

Na actualidade ambas ferramentas contemplan a análise de 10 substantivos en 3 linguas e aportan datos para 429 patróns sintácticos, 2536 estruturas sintáctico-semánticas (que resultan da interface entre a realización formal e a clasificación semántico-relacional), así como uns 2479 exemplos estándar. Un total de 60001 formas e 15718 lemas están adxudicados a diferentes clases semánticas.

[14] http://portlex.usc.gal/develop/lematiza
[15] http://portlex.usc.gal/develop/combina.php
[16] https://adimen.si.ehu.es/web/WordNet2TO
[17] http://wndomains.fbk.eu/
[18] http://www.adampease.org/OP/
[19] Os resultados da Figura 5 fan referencia á busca CQL en Sketch Engine `[lemma=''mudanza''] [lemma=''de''] [tag=''D.*'']? [tag=''A.*'']? [tag=''N.*''] .`
[20] API http://portlex.usc.gal/develop/es/api?ontology=top&category=Building e http://portlex.usc.gal/develop/es/api?ontology=epinonyms&category=ili-30-03259505-n&subcategories=on

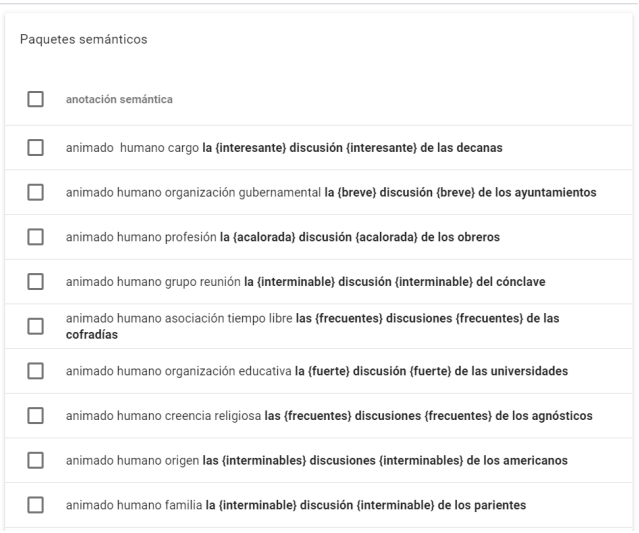

Figura 6: Clases semánticas para N1 de DISCUSIÓN

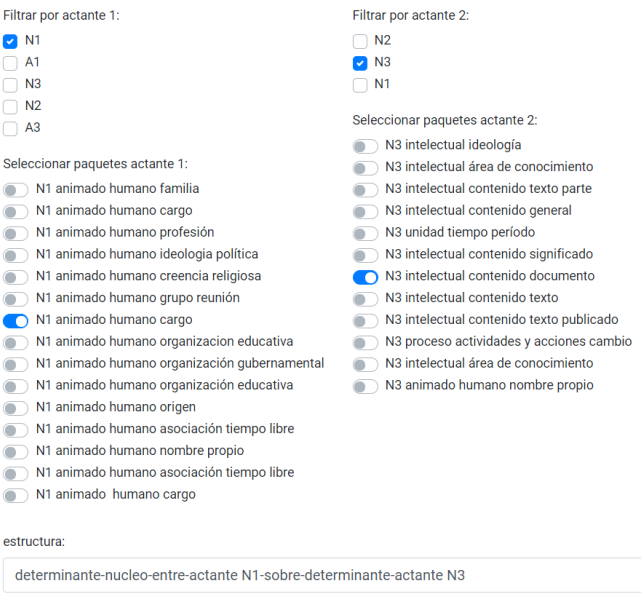

Figura 7: Interface de usuario en Combinatoria

Na primeira interface de acceso a **Xera**, o usuario selecciona nun despregable o idioma, o núcleo — neste caso DISCUSIÓN — e a estrutura argumental — N1: AXENTE ("Aquel ou aquilo que realiza unha acción") —, así como a clase ou clases semánticas prototípicas para o argumento concreto dun substantivo en cuestión (Figura 6)[21]. Cómpre engadir que os datos xerados seguen un principio de aleatoria predeterminada. Isto significa que as clases semánticas da cada argumento seguen un filtro semántico e a aleatoriedade atinxe aos representantes léxicos de cada clase, non ao rol semántico.

Un proceso semellante séguese na consulta do xerador de combinatoria biargumental, Combinatoria, que permite ao usuario seleccionar os actantes, paquetes e a combinatoria argumental (Figura 7), obtendo datos como os que recolle a Figura 8).

Porén, o tipo de datos envorcados non é a única diferenza entre ámbalas dúas ferramentas. **Combinatoria** tamén ofrece a posibilidade de cribar as combinatorias mediante o uso de *word embeddings*, representacións vectoriais dunha palabra en contexto, por tanto, un filtrado seguindo criterios de frecuencia de coaparición contextual (*vid.* parte superior da Figura 8). Para o desenvolvemento destes vectores recorremos ao método predictivo Word2vec (Mikolov et al., 2013), que fai uso dunha RNN (*Recurrent Neural Network*) de 2 niveis con dúas implementacións diferentes: o algoritmo Skip-gram tenta predicir o contexto máis adecuado dunha palabra analizando o seu vector e os vectores posteriores das palabras vinculadas ao vector da palabra orixe. O modelo CBOW tenta adiviñar a palabra máis adecuada para un contexto específico, é dicir, a palabra que máis frecuentemente ocupa un espazo: *a dor de* [espazo] *do neno*. Para a aplicación de Word2vec partimos dun modelo preadestrado de *Sketch Engine* e gardamos os resultados no formato propio de Word2vec (en matrices de N tamaño, onde N é o tamaño do vector preconfigurado para cada palabra). Para calcular a similitude entre os *tokens* calculamos a similitude entre os cosenos de dous vectores, tal e como representa a seguinte función matemática:

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|} \qquad (1)$$

O exemplo, tomado de (Bardanca Outeiriño, 2020), amosa así a distancia dos cosenos ou a similitude entre os lemas *rose* e *tulip* (Figura 9). Coa aplicación da fórmula a estes vectores obtemos un grao de similitude de 0.73:

$$\begin{aligned} a^T\cdot b &= 1\times1+1\times1+1\times1+1\times1+\\ &\quad 0\times1+1\times0+1\times0=4,\\ \|\vec{a}\| &= \sqrt{1^2+1^2+1^2+1^2+0^2+1^2+1^2}=2{,}44,\\ \|\vec{b}\| &= \sqrt{1^2+1^2+1^2+1^2+1^2+0^2+0^2}=2{,}23, \end{aligned}$$

$$similaridade = \frac{4}{2{,}44\times2{,}23} = 0{,}73$$

[21] Isto obsérvase comparando as clases semánticas da Figura 6 cos resultados dunha busca do argumento N3 ("AFECTADO: THEMA") coa mesma estrutura sintáctica [*determinante* + *adxectivo_o* + *discusión* + *adxectivo_o* + *de* + *determinante* + *actante:* N3]

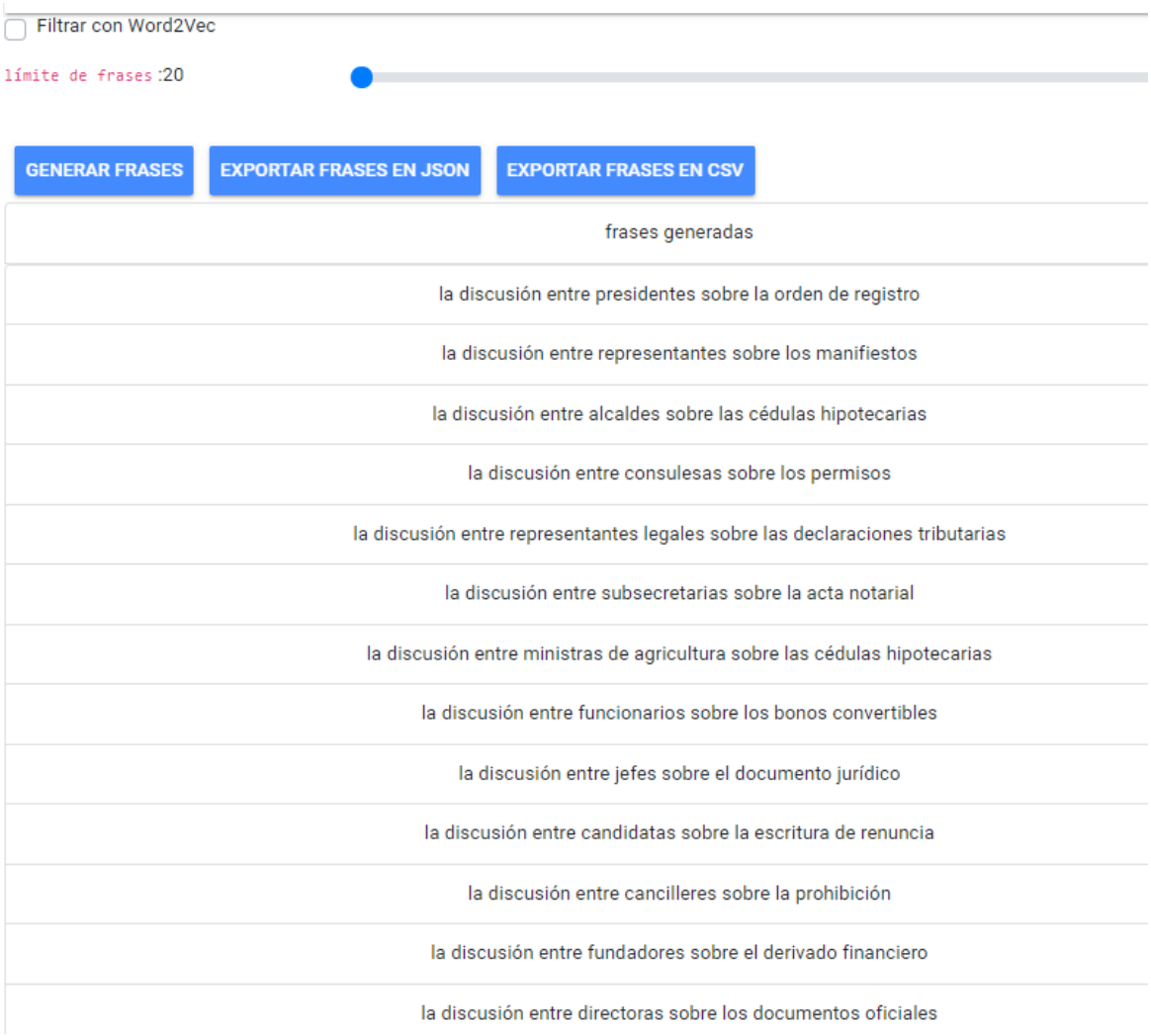

Figura 8: Envorcado de combinatoria

rose	0
carnation	1
peony	2
hydrangea	3
gerbera	4
flower	5
lily	6

$\vec{a}$

tulip	0	1	2	3	4

$\vec{b}$

rose	1	5	3	2	6

Figura 9: Similitude - Word2vec

Este mesmo procedemento é o que aplicamos para calcular a frecuencia de coaparición en contexto das palabras relacionadas cun *token*. Isto é, non usamos a análise vectorial para desambiguar significados, senón para filtrar os datos atendendo á compatibilidade semántico-contextual dos argumentos — ou sexa, a distribución no sentido de Firth (1957, p.11f) ou o significado combinatorio de Engel (2004, p.188). En definitiva, filtramos incompatibilidades do significado combinatorio, o que resulta especialmente relevante, xa que a corrección gramatical non implica directamente corrección ou adecuación semántico- comunicativa. Word2vec tamén se aplica na fase de contextualización (Figura 2) para cribar a coaparición en contexto da modificación adxectival (*vid.* 2.2).

4. Conclusións e traballos futuros

As principais dificultades no desenvolvemento dos prototipos foron as propias dos estudios multilingües e, en especial, a escolla da metodoloxía máis apropiada para a obtención e envorcado dos datos seguindo criterios non so formais, senón tamén semánticos. Establecer unha ontoloxía descriptiva que permitise combinar os trazos categoriais que servían de punto de partida — os da lexicografía valencial - coas ontoloxías de WordNet foi unha das tarefas máis laboriosas. Finalmente desenvolveuse unha ontoloxía propia seguindo unha aproximación *bottom up*.

Os prototipos xa están en funcionamento: da xeración de frases nominais simples — por tanto, frases con estrutura monoargumental correctas dende un punto de vista gramatical e aceptables semánticamente — encárgase o simulador *Xera*. *Combinatoria*, pola súa banda, aporta a xeración automática de frases nominais con estruturas biargumentais, tendo tamén entre as súas competencias a xeración do contexto frasal e do marco oracional. Perséguese, pois, o obxectivo de humanizar os resultados xerados. A curto prazo cómpre acometer:

1. tarefas de optimización dos propios recursos e a súa didactización, destacando aquí o aumento do número das unidades de análise así como de novos campos informativos. Queda por desenvolver o contexto oracional.

2. unha automatización dos procedementos analíticos. Hai que continuar traballando na aplicación combinada de ferramentas co fin de avanzar na automatización e optimización na extracción de información sintáctico-argumental e combinatoria. Nesta liña, uns dos aspectos que estamos a mellorar atinxe á extracción e tratamento do léxico expandido, que se viña facendo de xeito individualizado para cada unha das linguas e que supón, por exemplo, a reiteración de tarefas de depurado. Coa finalidade de evitar esta repetición de procedementos, desenvolveuse *TraduWord*, unha ferramenta de tradución do caudal léxico paradigmático a partir dos datos extraídos de WordNet. *TraduWord* xa foi aplicada no deseño dun prototipo de xeración automática para o galego e o portugués, *XeraWord*[22] (Bardanca Outeiriño et al., 2021), o cal se fundamenta na metodoloxía de *Xera* e *Combinatoria*. O deseño de ferramentas deste tipo abre, por tanto, unha nova canle de traballo.

Cara ao futuro poderíase explorar se a descripción da interface sintáctico-semántica mediante roles e clases semánticas, que da sustento aos xeradores, permitiría avanzar na obtención de resultados sistemáticos e regulares en prol da desambiguación de acepcións de significado. Algúns datos apuntan nesta dirección: así, no substantivo español DOLOR unha metaestructura do tipo [*de* + *determinante* + *nome*] é cómun á expresión do EXPERIMENTANTE (*el dolor del animal*), da ORIXE (*el dolor de la operación*) e da LOCACIÓN (*el dolor de espalda*). Porén, a aparición dun rol locativo e, por tanto, de clases semánticas como [animado humano órgano] ou [animado animal parte do corpo] é propia dunha acepción do tipo 'sensación molesta de tipo físico' e non dunha acepción relacionada co campo do sentimento. Por tanto, a análise dos exemplares léxicos xunto cos roles e clases semánticas podería ser un punto de partida. Ademais contamos con estudos preliminares sobre a modificación adxectival (López Iglesias, 2020) que permiten observar non só a especialización distribucional de determinados adxectivos, senón tamén a súa aparición preferente ou exclusiva segundo a acepción de significado.[23] Nesta liña, un estudo dos datos que envorca Word2vec (*vid.* 3) podería ser de interese para a análise da interface sintáctico-semántica.

[22] http://ilg.usc.gal/xeraword

Agradecementos

Esta investigación está en relación cos proxectos "Generación multilingüe de estructuras argumentales del sustantivo y automatización de extracción de datos sintáctico- semánticos" - MultiGenera (financiado pola Fundación BBVA) — Convocatoria de ayudas a equipos de investigación científica en Humanidades Digitales 2017), "Generador multilingüe de estructuras argumentales del sustantivo con aplicación en la producción en lenguas extranjeras" — MultiComb (financiado por FEDER/Ministerio de Economía, Industria y Competitividad — Agencia Estatal de investigación; 2018, FFI2017-82454-P) así como con "Ferramentas TraduWord e XeraWord: tradución de caudal léxico e xeración automática da linguaxe natural en galego e portugués" (2020-PU004, convocatoria de proxectos colaborativos, USC).

Referencias

Agirre, Eneko & Philip Edmonds. 2006. *Word sense disambiguation. algorithms and applications*. Dordrecht: Springer.

Álvez, Javier, Jordi Atserias, Jordi Carrera, Salvador Climent, Egoitz Laparra, Antoni Oliver & German Rigau. 2008. Complete and consistent annotation of wordnet using the top concept ontology. En *International Conference on Language Resources and Evaluation (LREC)*, 1529–1534.

Bardanca Outeiriño, Daniel. 2020. *Automatic generation of dictionaries*: Universidade de San-

[23] Así, por exemplo, para a expresión dunha dor de tipo físico son frecuentes en español adxectivos como *leve*, *persistente* ou *frecuente*, frente a *sincero*, *hondo* e *irreparable* que aparecen en realizacións do campo do sentimento.

tiago de Compostela. Trabajo de Fin de Máster.

Bardanca Outeiriño, Daniel, María Caíña Hurtado, María José Domínguez Vázquez, José Luis Iglesias Allones & Alberto Simões. 2021. Automatic generation of nominal phrases: Extending the tool *Xera* for portuguese and galician languages. No prelo.

Bentivogli, Luisa, Pamela Forner, Bernardo Magnini & Emanuele Pianta. 2004. Revising WordNet domains hierarchy: Semantics, coverage, and balancing. En *COLING Workshop on Multilingual Linguistic Resources*, 101–108.

Bond, Francis & Ryan Foster. 2013. Linking and extending an open multilingual WordNet. En *Meeting of the Association for Computational Linguistics (ACL)*, 1352–1362.

Delli Bovi, Claudio & Roberto Navigli. 2017. Multilingual semantic dictionaries for natural language processing: The case of BabelNet. En *Encyclopedia with Semantic Computing and Robotic Intelligence*, 149–163. World Scientific. 10.1142/9789813227927_0017.

Domínguez, María José, Miguel Anxo Solla & Carlos Valcárcel. 2019. Resources interoperability: exploiting lexicographic data to automatically generate dictionary examples. En *eLex Conference: Electronic lexicography in the 21st century*, 51–57.

Domínguez, María José & Carlos Valcárcel. 2020. PORTLEX as a multilingual and cross-lingual online dictionary. En *Studies on multilingual lexicography*, 135–158. 10.1515/9783110607659-008.

Engel, Ulrich. 2004. *Deutsche Grammatik – Neubearbeitung*. München: Iudicium.

Firth, John Rupert. 1957. *A synopsis of linguistic theory: 1930–1955*. Philological Society.

Gamallo, Pablo & Jose Ramom Pichel. 2007. Un método de extracción de equivalentes de traducción a partir de un corpus comparable castellano-gallego. *Procesamiento del Lenguaje Natural* 39. 241–248.

Geyken, Alexander, Frank Wiegand & Kay-Michael Würzner. 2017. On-the-fly generation of dictionary articles for the DWDS website. En *eLex Conference: Electronic lexicography in the 21st century*, 560–570.

Gómez Guinovart, Xavier & Miguel Anxo Solla. 2018. Building the Galician wordnet: methods and applications. *Language Resources and Evaluation* 52(1). 317–339. 10.1007/s10579-017-9408-5.

Hashimoto, Tatsunori, Hugh Zhang & Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. En *Conference of the North American Association for Computational Linguistics (NAACL)*, 1689–1701. 10.18653/v1/N19-1169.

Horacek, Helmut & Michael Zock. 2015. *New concepts in natural language generation: Planning, realization and systems*. London: Bloomsbury Academic.

Izquierdo, Rubén, Armando Suárez & German Rigau. 2007. Exploring the automatic selection of basic level concepts. En *International Conference on Recent Advances on Natural Language Processing (RANLP)*, 298–302.

Jiménez, Moreno, Luis Gil, Juan Manuel Torres-Moreno, Roseli S. Wedemann & Erich SanJuan. 2020. Generación automática de frases literarias. *Linguamática* 12(1). 15–30. 10.21814/lm.12.1.308.

Kabashi, Besim. 2018. A lexicon of Albanian for natural language processing. En *EURALEX International Congress: Lexicography in Global Contexts*, 855–862.

Kilgarriff, Adam, Milos Husák, Katy McAdam, Michael Rundell & Pavel Rychlý. 2008. GDEX: automatically finding good dictionary examples in a corpus. En *EURALEX International Congress*, 425–432.

Kosem, Iztok, Kristina Koppel, Tanara Zingano Kuhn, Jan Michelfeit & Carole Tiberius. 2019. Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography* 32. 119–137. 10.1093/ijl/ecy014.

López Iglesias, Nerea. 2020. *Analysing nominal phrase contexts for the automatic extraction of linguistic and lexicographic data*: Universidade do Minho. Trabajo de Fin de Máster.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. En *International Conference on Neural Information Processing Systems*, 3111–3119.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine J. Miller. 1990. WordNet: an on-line lexical database. *International Journal of Lexicography* 3(4). 235–244. 10.1093/ijl/3.4.235.

Nallapati, Ramesh, Bowen Zhou, Cicero dos Santos, Çağlar Gülçehre & Bing Xiang.

2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. En *Conference on Computational Natural Language Learning (CoNLL)*, 280–290. 10.18653/v1/K16-1028.

Navigli, Roberto & Simone Paolo Ponzeto. 2012. BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193. 217–250. 10.1016/j.artint.2012.07.001.

Niles, Ian & Adam Pease. 2001. Towards a standard upper ontology. En *International Conference on Formal Ontology in Information Systems*, 2–9. doi/10.1145/505168.505170.

Otter, Daniel. W., Julian R. Medina & Jugal K. Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems* 1–21. 10.1109/TNNLS.2020.2979670.

Padró, Lluís. 2012. Analizadores multilingües en FreeLing. *Linguamática* 3(2). 13–20.

Renau, Irene & Rogelio Nazar. 2016. Automatic extraction of lexical patterns from corpora. En *EURALEX International Congress: Lexicography and Linguistic Diversity*, 823–830.

Roemmele, Melissa. 2016. Writing stories with help from recurrent neural networks. En *Conference on Artificial Intelligence (AAAI)*, 4311–4312.

Sagot, Benoît & Darja Fišer. 2008. Building a free French wordnet from multilingual resources. En *OntoLex 2008 Workshop*, s/pp.

Sordoni, Alessandro, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao & Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. En *Annual Conference of the North American Chapter of the ACL*, 196–205. 10.3115/v1/N15-1020.

Trap-Jensen, Lars. 2018. Lexicography between nlp and linguistics: Aspects of theory and practice. En *EURALEX International Congress: Lexicography in Global Contexts*, 25–37.

Vicente, Marta, Cristina Barros, Francisco Agulló, Fernand S. Peregrino & Elena Lloret. 2015. La generación el lenguaje natural: análisis del estado actual. *Computación y Sistemas* 19(2). 721–756. 10.13053/CyS-19-4-2196.