Multilingual automatic generation of noun valency patterns and their sentence contexts

María José Domínguez Vázquez, Carlos Valcárcel Riveiro
Universidade de Santiago de Compostela - ILG, Universidade de Vigo

Fundación BBVA · Universidade de Vigo · USC Universidade de Santiago de Compostela · Instituto da Lingua Galega (ILG) · Gobierno de España · Ministerio de Economía, Industria y Competitividad · Unión Europea · Fondo Europeo de Desarrollo Regional "Una manera de hacer Europa"

2012 PORTLEX → 2018 MultiGenera → 2020 XERAWORD → 2021 MultiComb

Portlex: The dictionary Portlex, the starting point, is a multilingual, digital, semi collaborative and cross-lingual resource for the valency of the nominal phrase in five languages.

MultiGenera: Development of a simulator for the automatic generation of noun phrases in Spanish, German and French.

MultiComb: Automatic generation of the phrasal and sentence contexts for the previously selected nouns in MultiGenera

XeraWord e TraduWord: Development of the MultiGenera prototype for Galician and Portuguese. Automatic translation of lexical packages.

Key word of the projects: Resource Interoperability

METHODOLOGICAL APPROACHES

VALENCY, SEMANTIC FEATURES, AND SURFACE REALIZATION PATTERNS

DESCRIPTION OF REALIZATION PATTERNS FOR ARGUMENT STRUCTURES

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

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

ACKNOWLEDGEMENTS

The results of this work are related to the research project "Generación multilingüe de estructuras argumentales del sustantivo y automatización de extracción de datos sintáctico-semánticos", financed by the BBVA Foundation Grants for Scientific Research Teams 2017, to the research project "Multilingual generator of noun argument structures with application in foreign language production", financed by the spanish Ministry of Economy, Industry and Competitiveness (Scientific and technical excellence research program, FFI2017-82454-P) and to the research project "*TraduWord* and *XeraWord* Tools: lexical units translation and automatic language generation for Galician and Portuguese", financed by University of Santiago de Compostela (2020).

FURTHER INFORMATION
http://portlex.usc.gal/
https://ilg.usc.es/

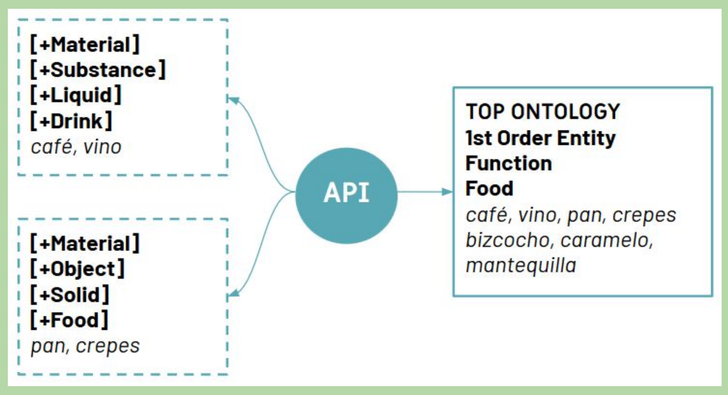
Workshop Red INTELE, 25/11/2020

EXPANSION OF LEXICAL PROTOTYPES

For each argument-role-slot a general list of prototypical lexical items will be obtained, as shown here for the Spanish argument structure 'olor a' + *common noun*; each item is then annotated with semantic features. Corpus query tool: SketchEngine

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

GENERATION OF ARGUMENT STRUCTURE SURFACE REALIZATIONS



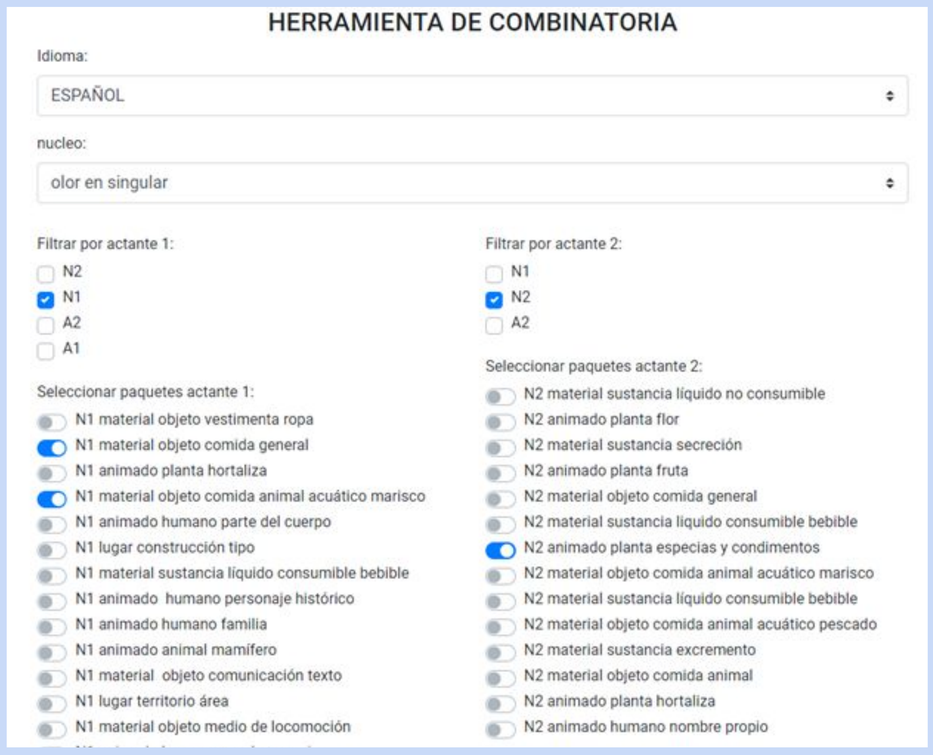
For the automatic generation of the argument structures, we use own python scripts and an own API for accessing wordnets and semantic ontologies. The lists of candidate lexical items to fill in each argument slot will allow users of our tool to choose those they prefer to generate simple noun phrases.

AUTOMATIC GENERATION

PATTERNS

Screenshot of the interface for biargumental pattern generation.

Word2vec was also implemented to further improve the generation biargumental structures.



CONTEXT GENERATION

At phrasal level. In order to obtain a more varied and human-like output, word2vec is also used and some studies on lexical functions (Mel'čuk 2013, 2015) have been carried out. Examples:

un **fuerte** **olor** *a tabaco* ('a strong smell of tobacco'),
aquel **agradable** **olor** *a madera de su habitación* ('that pleasant smell of wood in his room'),

At sentence level. The first step is the corpus data extraction following frequency criteria. For this purpose a custom-made Post-Tagger was developed. The aim is to generate sentence contexts embedding the most frequent valency patterns. Examples:

Subject (NP) + Copula + Attribute.
El **olor** *a tabaco de la casa* **resultaba** **insoportable**.
'*The tobacco smell in the house was unbearable*'.

Subject + Verb + Object (NP).
El **vecindario** **sentía** *el* **olor** *a tabaco de la casa.*
'*the neighborhood noticed the tobacco smell of the house*'.