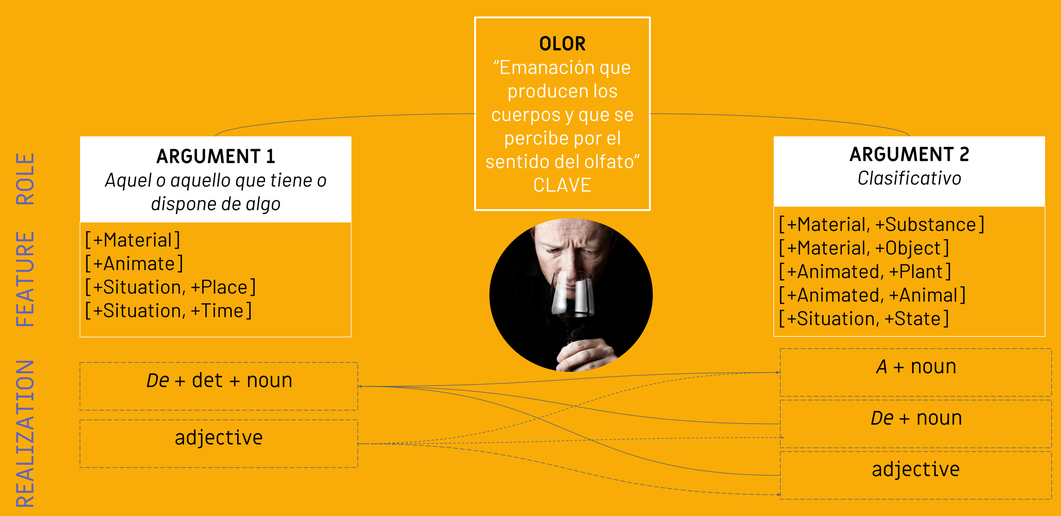
Multilingual generation of noun valency patterns for extracting syntactic-semantical knowledge from corpora

María José Domínguez Vázquez, Carlos Valcárcel Riveiro, David Lindemann
Universidade de Santiago de Compostela, Universidade de Vigo, Universität Hildesheim

Fundación BBVA
GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD
FONDO EUROPEO DE DESENVOLVEMENTO REXIONAL "Unha maneira de facer Europa" UNIÓN EUROPEA

2012 PORTLEX → 2018 MultiGenera → 2021 MultiComb

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*.

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

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, and 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).

FURTHER INFORMATION
portlex.usc.gal/multigenera

PROJECT GOALS: MULTIGENERA

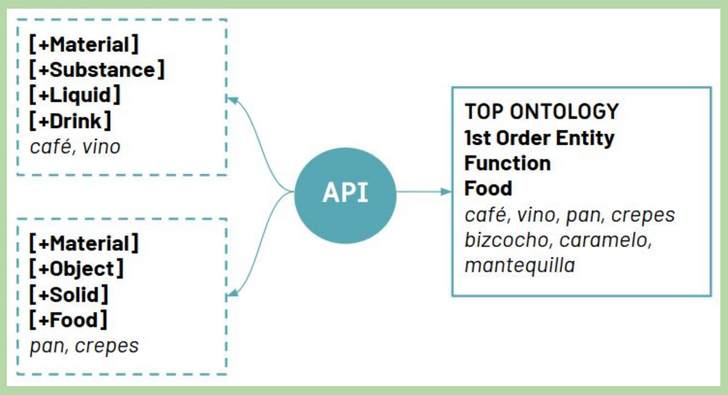
The **aim of the project** is to develop a prototype for a generator of argument structure or valency realisations in terms of syntagmatic and paradigmatic combinations of Spanish, German and French nouns.

The **two main applications** of the tool prototype:

(1) the generation of noun phrases as argument structure realizations that follow patterns related to semantic features, for the creation of corpus and web query strings, and

(2) the knowledge-based generation of simple and complex noun phrases that are acceptable in a coherent sentence context.

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.

OUTLOOK: CONTEXT GENERATION

MultiComb

For context generation at ph For a formal modelling and machine-readable annotation of structures like i.e. in Spanish *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'), a selection of basic lexical functions (LF) is carried out, following Mel'čuk (2013, 2015). The paradigmatic sets associated to LF will depend not only on each noun, but also on the specific lexical restrictions of each of the three languages. For example, in the case of the Spanish noun *olor* 'smell', we would obtain the following prototype lists for the selected LF:

Magn (olor) = fuerte ('strong'), intenso ('intense'), penetrante ('pungent, penetrating')
AntiBon (olor) = malo ('bad'), desagradable ('unpleasant'), nauseabundo ('nauseating'), rancio, ('rancid'), insoportable ('unbearable'), asqueroso ('nasty'), fétido ('foul')
Bon (olor) = agradable ('pleasant'), fresco ('fresh'), dulce ('sweet')
Ver (olor) = característico ('characteristic'), genuino ('genuine'), verdadero ('real')

This allows to program randomly the appearance of adjectives linked to a noun by a LF, and obtain a more varied and human-like output.

CONTEXT GENERATION AT SENTENCE LEVEL

In this stage, the noun phrases (Det + noun + arguments) previously generated will fill in the valency slots of a verb. These sentence contexts will be limited to four basic syntactic structures: [Subject (NP) + Verb: *el olor a tabaco de la casa se disipó*, 'the tobacco smell in the house faded away'], [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'] and [Subject+Verb+Prepositional Complement (Prep + NP): *Me enamoré del olor a campo de su ropa*, 'I fell in love with the country smell of their clothes']. This will allow to generate sentence contexts with the most frequent valency patterns. Again, new sets of lexical prototypes will be created for the rest of slots of the sentence contexts on the basis of frequency queries in corpora and dictionaries.

MultiGenera Scientific Committee