



HERRAMIENTAS Y DIFICULTADES EN EL ANÁLISIS DEL GRUPO NOMINAL EN FRANCÉS PARA SU PROCESAMIENTO COMPUTACIONAL

TOOLS AND DIFFICULTIES IN THE ANALYSIS OF THE FRENCH NOUN PHRASE FOR COMPUTATIONAL PROCESSING

Carlos Valcárcel Riveiro
Universidade de Vigo
carlos.valcarcel@uvigo.gal

Laura Pino Serrano
Universidade de Santiago de Compostela
laura.pino@usc.es

RESUMEN

La presente contribución presenta el trabajo desarrollado sobre la lengua francesa en el portal lexicográfico multilingüe PORTLEX. Concretamente, el artículo se centra en las herramientas utilizadas por los equipos de francés de tres proyectos de investigación: el diccionario PORTLEX y el desarrollo de prototipos para la generación automática de sintagmas nominales en los proyectos *MultiGenera* y *MultiComb*. Se detallan los recursos utilizados distinguiendo las herramientas de consulta (diccionarios, gramáticas) de las computacionales (corpus, bases de datos). Asimismo, se describen las principales dificultades asociadas al uso de cada herramienta o recurso y se explican las decisiones adoptadas para solventarlas o minimizarlas.

Palabras clave: francés, herramientas, recursos, lexicografía, PLN.

ABSTRACT

This paper presents the work carried out on the French language within the multilingual lexicographic portal PORTLEX. In particular, the article focuses on the tools used by the teams working on French in three research projects: the PORTLEX dictionary and the development of prototypes for the automatic generation of noun phrases within *MultiGenera* and *MultiComb* projects. The resources used are described in detail, distinguishing between reference tools (dictionaries, grammars) and computational tools (corpora, databases). In addition, the main difficulties associated with the use of each tool or resource are described and the choices made to resolve or minimise them are explained.

Keywords: French, tools, resources, lexicography, NLP.



1. INTRODUCCIÓN

Desde sus inicios, hace ya diez años, el francés ha sido una de las lenguas de trabajo en el portal lexicográfico multilingüe PORTLEX¹. Este portal se creó en la Universidad de Santiago bajo la dirección de Domínguez Vázquez durante el desarrollo del diccionario del mismo nombre. Más allá de un espacio web centrado en la lexicografía, PORTLEX constituye un lugar de encuentro y colaboración de especialistas en los ámbitos de la lingüística y de la computación para desarrollar proyectos de investigación, herramientas lingüísticas y diferentes acciones de divulgación científica (congresos, cursos, talleres, etc.). Por lo tanto, la inclusión del francés en las actividades de este portal, y toda la investigación paralela desarrollada en diferentes áreas de la lingüística, ha implicado la colaboración de diferentes especialistas en esta lengua. Así, a lo largo de estos años grupos de francesistas, en colaboración con otros equipos especializados en las otras lenguas de trabajo, han realizado diferentes tareas de tipo teórico y aplicado: el análisis y anotación sintáctico-semántica de frases nominales, la identificación y anotación semántica de prototipos léxicos, la elaboración de paquetes léxicos, la revisión de textos generados automáticamente o incluso el diseño de bases de datos lingüísticos, entre otras.

En la presente contribución trataremos, pues, de la labor desarrollada por estos equipos de francés en PORTLEX centrándonos en las herramientas utilizadas y las dificultades encontradas a la hora de afrontar todas estas tareas. Como era de esperar, estas últimas fueron numerosas y de diverso tipo. Si bien una parte de estos desafíos eran comunes a las otras lenguas meta de los proyectos, varios se relacionaban directamente con particularidades de la lengua francesa. En primer lugar, describiremos las tareas realizadas y los resultados obtenidos en los tres proyectos desarrollados en el portal hasta el

¹ <http://portlex.usc.gal/>

momento: el diccionario PORTLEX (FFI2012-32456)² (Domínguez Vázquez & Valcárcel Riveiro, 2020) y los prototipos de generación automática de frases nominales de los proyectos *MultiGenera*³ y *MultiComb* (FFI2017-82454-P)⁴ (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019; Domínguez Vázquez, Bardanca Outeiriño & Simões, 2021). Haremos también referencia al etiquetador semántico desarrollado en el proyecto en curso ESMAS-ES⁺ (PID2022-137170OB-I00)⁵. Pasaremos después a hablar de las principales herramientas de trabajo del equipo de francés de PORTLEX, haciendo una distinción entre herramientas de consulta (diccionarios y gramáticas) y herramientas de procesamiento del lenguaje natural (corpus, bases de datos y bibliotecas). Describiremos sus funcionalidades y limitaciones más destacadas, así como las dificultades que estas últimas han supuesto para el trabajo con la lengua francesa. En las conclusiones extraeremos las principales lecciones aprendidas y mencionaremos algunas líneas de trabajo futuras en el marco del portal PORTLEX.

2. LOS PROYECTOS PORTLEX Y SUS APORTACIONES AL FRANCÉS

Los cuatro proyectos desarrollados en el marco del portal lexicográfico PORTLEX tienen en común tres aspectos esenciales: su carácter multilingüe, su orientación dependencial y su interés por el grupo nominal. En primer

² El proyecto *Portal Lexicográfico: Diccionario online modular multilingüe y corpus informatizado anotado de la frase nominal* fue financiado por el Ministerio de Economía y Competitividad, por la Unión Europea a través del Fondo Europeo de Desarrollo Regional (FEDER) 2007-2013 y por la Red de Lexicografía RELEX (R2014/042), 2013-2015.

³ Este proyecto fue financiado por la Fundación BBVA a través del programa de Ayudas Fundación BBVA a Equipos de Investigación Científica - Humanidades Digitales, 2017-2020.

⁴ Este proyecto fue financiado por la Agencia Estatal de Investigación (Ministerio de Ciencia e Innovación) en el marco del Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia, Generación de Conocimiento (EXCELENCIA 2017, 2017-PN091) y por la Unión Europea a través del Fondo Europeo de Desarrollo Regional (FEDER) “Una manera de hacer Europa”, 2018-2021.

⁵ Este proyecto está financiado por la Agencia Estatal de Investigación (Ministerio de Ciencia e Innovación) y por la Unión Europea a través del Fondo Europeo de Desarrollo Regional (FEDER) “Una manera de hacer Europa”.

lugar, la inclusión de múltiples idiomas en el análisis lingüístico permite una comprensión más amplia y comparativa de las estructuras lingüísticas. Así, al abordar diferentes idiomas, los proyectos PORTLEX pueden identificar similitudes y diferencias en la construcción de los grupos nominales y sus combinatorias en diversas lenguas. De hecho, el desarrollo de interfaces que simplifiquen la consulta contrastiva de la información de varias lenguas ha sido una preocupación de los equipos de PORTLEX desde sus inicios. Así, en los recursos disponibles en este portal lexicográfico, los datos referentes al francés pueden ser contrastados de diferentes formas con los de las otras lenguas de trabajo, principalmente el castellano y el alemán.

En segundo lugar, el enfoque valencial o dependencial aplicado al estudio del grupo nominal tampoco deja de constituir un aspecto novedoso en los estudios franceses. A pesar de que la gramática dependencial moderna fue formulada inicialmente por un lingüista francés, Tesnière (1959), los análisis de corte valencial todavía son menos frecuentes para el francés que para otras lenguas como el alemán. Menos aún lo son las investigaciones sobre los predicados nominales, ya que el análisis del predicado verbal y de sus argumentos ha ocupado siempre un lugar prioritario en el campo de la gramática de dependencias. Sin embargo, este verbocentrismo no ha frenado el desarrollo de investigaciones sobre los argumentos del sustantivo, su combinatoria y las restricciones formales y semánticas asociadas. Para el francés destacan en esta línea los estudios desarrollados en el marco de la Teoría Sentido-Texto (Mel'čuk, 1997) y del *Laboratoire d'Informatique Documentaire et Linguistique* (LIDL) (Gross, 2012). Como veremos más adelante, estos y otros estudios resultan, pues, esenciales para el trabajo desarrollado en el portal PORTLEX sobre la lengua francesa.

2.1. *EL DICCIONARIO PORTLEX*

La elaboración de este diccionario electrónico constituyó el primer proyecto del portal y se desarrolló entre los años 2013 y 2017. Esencialmente, PORTLEX es un diccionario multilingüe y valencial sobre la estructura del

grupo nominal en cinco idiomas, entre ellos el francés. Por lo tanto, las entradas están constituidas por lexemas nominales y en ellas se analizan sus diferentes argumentos, así como la combinatoria que se da entre ellos. Probablemente, los aspectos más novedosos de este diccionario son la estructura de la base de datos que lo sustenta y la interfaz de consulta contrastiva, que permite visualizar en pantalla entradas equivalentes en dos lenguas (Domínguez Vázquez & Valcárcel Riveiro, 2020).

Para la lengua francesa se elaboraron un total de 37 entradas, lo que supuso la descripción de 197 realizaciones actanciales⁶ y 152 combinatorias. En ambos casos se proporciona un análisis pormenorizado en el que se detalla la función sintáctica de las realizaciones, su rol semántico, los rasgos semánticos asociados y una descripción de su estructura formal. Además, tanto las realizaciones como las combinatorias se ilustran con un mínimo de tres ejemplos buscados y seleccionados manualmente en corpus. Para el francés se extrajeron un total 1047 ejemplos y, sin duda, esta fue la parte más laboriosa del proyecto. Esto se debió principalmente a las limitaciones del corpus utilizado que se comentarán más adelante. Prácticamente todo el trabajo sobre el francés en el diccionario PORTLEX corrió a cargo de Valcárcel Riveiro, de la Universidad de Vigo⁷.

2.2. MULTIGENERA Y MULTICOMB

Como ya se ha comentado, una de las mayores dificultades experimentadas por los equipos de trabajo durante la elaboración del diccionario PORTLEX fue la búsqueda y selección de ejemplos válidos. Para intentar solventar este problema se desarrollaron entre 2018 y 2021 dos proyectos paralelos de

⁶En lo referente a las realizaciones actanciales analizadas para el francés, cabe destacar como novedad que el diccionario incluya las estructuras apositivas N_1N_2 (p. ej., *la consommation poisson, une question santé*) (Valcárcel Riveiro, 2017).

⁷ Como parte de su programa de formación, varios estudiantes del Máster europeo de lexicografía EMLex editaron y revisaron algunas entradas. Se trató en concreto de Océane Meyan, Nikolay Chepurnykh y Polina Mikhel.

investigación aplicada que tenían como objetivo el desarrollo de dos prototipos para generar, en una primera fase, sintagmas nominales (*Combinatoria*)⁸ y, en una segunda, contextos oracionales para estos (*CombiContext*)⁹. Aun tratándose de prototipos, nos encontramos ante herramientas complejas, ya no solo por su carácter multilingüe, sino sobre todo porque son los propios usuarios quienes establecen los parámetros de los sintagmas y oraciones generados en función de sus necesidades. Las estructuras creadas por estos prototipos pueden servir para diversos fines, pero principalmente para nutrir con ejemplos personalizados herramientas lexicográficas (diccionarios, glosarios) y actividades de aprendizaje de lenguas (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019; Domínguez Vázquez, Bardanca Outeiriño & Simões, 2021).

Los prototipos desarrollados en *MultiGenera* y *MultiComb* cubren tres lenguas (alemán, castellano y francés) y veinte lexemas o núcleos nominales. Además de las complicaciones computacionales, la construcción de estos prototipos requirió un ingente trabajo a nivel lingüístico por parte de los equipos investigadores involucrados: identificación de prototipos léxicos para cada realización argumental, etiquetado semántico de estos prototipos¹⁰, creación y depuración de paquetes semánticos para generar argumentos nominales y validación tanto de combinatorias de realizaciones como de estructuras oracionales (Domínguez Vázquez, Valcárcel Riveiro & Lindemann, 2018; Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019; Domínguez Vázquez, Bardanca Outeiriño & Simões, 2021). Asumir toda esta carga de trabajo requirió conformar un equipo más consolidado para la lengua francesa. Además de Valcárcel Riveiro, que asumió también tareas computacionales, trabajaron en estos proyectos Pino Serrano (Universidade de Santiago de Compostela),

⁸ <http://portlex.usc.gal/combinatoria/usuario>

⁹ <http://portlex.usc.gal/combinatoria/verbal>

¹⁰ Para tal fin se elaboró una ontología léxica propia (Domínguez Vázquez, Valcárcel Riveiro & Bardanca Outeiriño, 2021).

centrada en las tareas de análisis gramatical y validación de estructuras, y Malingret (Universidade de Santiago de Compostela), encargada de la revisión y evaluación de las estructuras generadas por los prototipos.

En total, trabajamos con 20 lexemas nominales¹¹ de diferentes campos semánticos (*absence, amour, augmentation, conversation, couleur, déménagement, discussion, douleur, fuite, largeur, mort, odeur, présence, question, réponse, saveur, séjour, texte, vidéo, voyage*), para los que se procesaron 152 realizaciones. Por parte del equipo esto implicó la elaboración y revisión de 624 paquetes semánticos para generar automáticamente estas realizaciones, la validación de 1491 estructuras oracionales y la revisión sistemática de los sintagmas y oraciones generados. Si bien los prototipos desarrollados ofrecen una cobertura muy limitada, la gran cantidad de datos analizados sin duda ha permitido un mejor conocimiento sobre el funcionamiento del grupo nominal en francés. Además, se han desarrollado herramientas específicas (API, Combina, Lematiza) para facilitar la extracción y procesamiento de datos léxicos en francés y otros idiomas (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019). Finalmente, todo esto nos ha permitido implementar una metodología para analizar en detalle las características semánticas de los argumentos nominales (Valcárcel Riveiro & Pino Serrano, 2023).

2.3. ESMAS-ES+

Actualmente, dentro del portal PORTLEX se está desarrollando ya un nuevo proyecto en el que también trabaja un equipo de francesistas: ESMAS-ES+. Además de Valcárcel Riveiro, componen el equipo las profesoras Vázquez Rodríguez y Castro Lorenzo, ambas de la Universidad de Vigo. Con este nuevo proyecto se pretende construir un prototipo de etiquetador semántico multilingüe y sostenible, es decir, que aproveche los recursos elaborados

¹¹ Es decir, solo se trabajó con una determinada acepción de estos vocablos, no con todas. Por ejemplo, para *question* los prototipos cubren la acepción ‘pregunta dirigida a alguien’ y no otras como ‘tema’ o ‘problema’.

en el marco de PORTLEX en proyectos anteriores. Principalmente, se parte de la ontología y los paquetes semánticos desarrollados para *MultiGenera* y *MultiComb* para alimentar el etiquetador semántico. Esto conlleva la revisión y traducción¹² de los paquetes de datos existentes, lo que por ahora limita la cobertura del etiquetador a los sustantivos. Además, se prevé el desarrollo de una base de datos y una interfaz específicas.

3. LAS HERRAMIENTAS DE TRABAJO: VENTAJAS, LIMITACIONES Y DIFICULTADES

Tanto en los proyectos *MultiGenera* y *MultiComb*, como en la elaboración del diccionario PORTLEX, el equipo de francés tuvo que emplear diversas herramientas a lo largo de las distintas fases de trabajo. Por un lado, el análisis semántico y sintáctico de las frases nominales, esencial en los tres proyectos realizados, demandaba el uso de herramientas de consulta, como diccionarios y gramáticas. Por otro lado, la extracción y procesamiento computacional de la gran cantidad de datos manejados exigía la búsqueda y manejo de corpus, repositorios y bases de datos disponibles para la lengua francesa, así como el desarrollo de herramientas específicas. En este apartado, describiremos las principales herramientas utilizadas para ambos fines, deteniéndonos en las dificultades encontradas por nuestro equipo.

3.1. HERRAMIENTAS DE CONSULTA

En todos los proyectos desarrollados en el marco de PORTLEX el equipo de lingüistas se encontró ante el desafío de analizar numerosas estructuras argumentales de sustantivos en las diferentes lenguas de trabajo. En el caso del francés, los predicados nominales estudiados a diferentes niveles se acercan ya a la cincuentena. Para llevar a cabo este análisis previo se tornó esencial contar con obras de consulta fiables y orientadas al tipo de marco

¹² En los proyectos anteriores no se elaboraron los mismos paquetes semánticos para todas las lenguas. Para traducir los paquetes de una lengua a otra se desarrolló otra herramienta, TraduWord, en colaboración con el Instituto da Lingua Galega (Domínguez Vázquez, Baranca Outeiriño & Simões, 2021, p. 278).

conceptual en el que se desarrollaban los proyectos: la gramática valencial o de dependencias. Estas fuentes no solo brindaron respuestas a ciertos interrogantes que fueron surgiendo durante el desarrollo de los diferentes proyectos, sino que permitieron sobre todo visualizar modelos para analizar esquemas argumentales de sustantivos, proporcionando así una base sólida para la comprensión y generación automática de las estructuras lingüísticas estudiadas.

3.1.1. *Diccionarios y léxicos*

Todo procesamiento automático del lenguaje implica ineludiblemente trabajar con el léxico y, por lo tanto, la disponibilidad de recursos lexicográficos es indispensable en la mayoría de los proyectos. Las necesidades lexicográficas de cada proyecto varían considerablemente en función de su orientación monolingüe o multilingüe, pudiendo incluir diccionarios generales y especializados, así como léxicos técnicos y bases terminológicas. En el caso de los proyectos desarrollados en el marco de PORTLEX se requerían recursos lexicográficos con amplia información sobre predicados verbales y, sobre todo, nominales. Desgraciadamente, la disponibilidad de diccionarios o léxicos de orientación valencial es escasa y a menudo los recursos existentes no tienen una gran cobertura. El francés no constituye una excepción, pero cuenta con un diccionario electrónico de amplia cobertura y de libre acceso que responde a las necesidades de los proyectos desarrollados: el *Trésor de la langue française informatisé* o *TLFi* (Dendien, 2020).

Si bien el *TLFi* no podría definirse como un diccionario valencial en sentido estricto (Domínguez Vázquez, 2018), su amplia cobertura y sobre todo la información específica que ofrecen sus entradas sobre cuestiones gramaticales lo convierten en una obra de referencia indispensable. Sin duda, se trata de una obra monumental que contiene más de 100.000 entradas y 270.000 definiciones. Incluye numerosos elementos metatextuales como códigos gramaticales, etiquetas semánticas y estilísticas, así como indicadores de dominio. Además de esto, se pueden encontrar 430.000 ejemplos extraídos de

obras literarias. Finalmente, *TLFi* también ofrece información detallada de carácter etimológico y bibliográfico (Pierrel, 2003, pp. 159-161).

Sin embargo, a pesar de toda la información compilada en el *TFLi*, este diccionario no explicita en las entradas de muchos sustantivos los complementos esenciales o argumentos. En muchos casos, y no de manera sistemática, los argumentos vienen expresados en la propia definición de los lemas. Por ejemplo, en el caso de la primera acepción de *séjour*, con la que trabajamos en MultiComb, el *TFLi* la define como “Fait de demeurer un certain temps dans un lieu, un endroit”. Esta definición evoca el argumento ‘agente’ (*le séjour de Marie*) y el argumento locativo (*le séjour en Italie*) como parte del significado, es decir, del predicado de *séjour* (*le séjour de Marie en Italie*). En otros casos, los argumentos se expresan mediante ejemplos en la sección SYNT. (*syntagme, syntaxe*) mezclados con complementos no esenciales. Así, en el caso de la primera acepción de *odeur*¹³ argumentos como *l’odeur du bois* o *odeur d’herbe* aparecen en la misma sección SYNT. que *odeur forte* o *odeur délicieuse*. Esta falta de coherencia en la presentación de la información sintáctica de los sustantivos hace a veces difícil la consulta de este recurso tan importante.

Otro recurso lexicográfico esencial para el equipo de francés en los proyectos PORTLEX es el *Dictionnaire explicatif et combinatoire du français contemporain (DECFC)* (Mel’čuk, Arbatchewsky-Jumarie, Iordanskaja, Mantha & Polguère 1984-1999). Los cuatro volúmenes publicados de esta obra tienen como objetivo proporcionar una descripción completa y sistemática del léxico francés siguiendo los principios de la lexicología explicativa y combinatoria (Mel’čuk, Clas & Polguère, 1995). Desarrollado en el marco de la Teoría del Sentido-Texto (TST), el DECFC pretende presentar, de forma coherente y sistemática, toda la información necesaria para que los hablantes

¹³ La definición ofrecida por el TLFi es la siguiente: “Émanation propre à un corps pouvant être perçue par l’homme ou par un être animé grâce à des organes particuliers et avec des impressions diverses (agréable, désagréable, indifférente)”.

expresen de forma lingüísticamente correcta cualquier significado que deseen comunicar. Para ello, incluye detalles exhaustivos para cada unidad léxica, como pronunciación, categoría gramatical y patrones sintácticos, así como las propiedades y restricciones de las coocurrencias léxicas y la combinatoria argumental (Meřćuk & Milićević, 2014, pp. 269-290). La marcada orientación valencial de este recurso, que incluye el análisis minucioso de los esquemas argumentales de numerosos sustantivos, hacen del *DECFC* una obra de consulta obligada para el equipo de francés del portal PORTLEX. Sin embargo, la cobertura de este diccionario es muy limitada ya que sólo describe 510 vocablos en francés y, salvo *conseil*, *apprentissage*, *maladie* y *risque*, analizados parcialmente en el diccionario PORTLEX¹⁴, el resto de sustantivos cubiertos por los diferentes proyectos no se encuentran en el *DECFC*. De todas formas, este recurso constituye un referente para el análisis de predicados nominales y de las combinatorias que se dan en ellos.

Además del *DECFC*, dos recursos desarrollados en el marco de la TST resultaron de gran utilidad en momentos puntuales del trabajo. Por un lado, consultamos el *Lexique actif du français (LAF)* (Meřćuk & Polguère, 2007), una versión accesible del *DECFC* y con una marcada orientación didáctica. De hecho, este recurso está dirigido tanto a profesorado y aprendientes del francés como a traductores y otros profesionales del lenguaje. De nuevo, la cobertura, aunque es más amplia que la del *DECFC*, sigue siendo insuficiente, ya que se analiza un número limitado de lexemas (781), agrupados en 386 vocablos. Por otro lado, y aunque se trata de un diccionario de lengua española centrado en las colocaciones, el *Diccionario de colocaciones del español (DiCE)* (Alonso-Ramos, 2004) se centra en el análisis de sustantivos, para los que se describen de manera sistemática los esquemas argumentales de sus

¹⁴ Cabe señalar aquí que, mientras el *DECFC* describe vocablos y sus diferentes lexemas o acepciones, el diccionario PORTLEX se limita normalmente a presentar información sobre un lexema o acepción específicos.

predicados (Alonso-Ramos, 2017, pp. 185-192). Dada la proximidad tipológica del francés y el castellano, resultó de interés consultar el análisis que presenta de lexemas equivalentes a los analizados en los proyectos PORTLEX (p. ej., *texto*, *discusión*) o pertenecientes a su mismo campo semántico.

3.1.2. Gramáticas e investigaciones gramaticales

Dada la fuerte orientación gramatical de los proyectos desarrollados en el portal PORTLEX, la consulta de investigaciones en este campo resultó imprescindible. No sólo fue necesario verificar la gramaticalidad de numerosas construcciones, sino también clarificar cuestiones conceptuales más básicas como, por ejemplo, determinar el carácter valencial de las realizaciones adjetivales de ciertos argumentos¹⁵. Como ya se ha indicado, el enfoque adoptado para el análisis gramatical en los proyectos PORTLEX es el dependencial o valencial, más concretamente el elaborado por Domínguez Vázquez (2011) para los grupos nominales del castellano y el alemán a partir de la gramática de Engel (2004). Sin embargo, el trabajo con la lengua francesa suscita cuestiones particulares a esta lengua que requieren el manejo de obras específicas.

Naturalmente, se consultaron gramáticas de referencia para el francés, entre las que cabe destacar el clásico *Le bon usage* (Grevisse & Goosse, 2008). Resultó de especial interés la reciente *Grande grammaire du français* (Abeillé & Godard, 2021) por su tratamiento del grupo nominal más próximo al enfoque de los proyectos PORTLEX. Por otro lado, también se consultaron con frecuencia algunas gramáticas de autor como la *Grammaire méthodique du français* (Riegel, Pellat & Rioul, 2009) y, más concretamente, la *Grammaire critique du français de Wilmet* (1997) debido a su estudio exhaustivo

¹⁵ Así, si en el grupo nominal *le séjour italien de tes amis* el carácter valencial del adjetivo *italien* (= *en Italie*) parece claro, ya no lo parece tanto en el syntagma *un texte italien sur les émotions*. Las investigaciones revelan que un mismo adjetivo puede ocupar una casilla valencial o circunstancial dependiendo del predicado, es decir, del contexto (Rigau, 1999).

de los sintagmas nominales, especialmente de la determinación. Finalmente, la marcada orientación semántica de la *Grammaire du sens et de l'expression* de Charaudeau (1992) resultó particularmente útil en el análisis de los roles semánticos de los argumentos nominales.

Más allá de todas estas obras de referencia, dos escuelas gramaticales de corte dependencial constituyeron los principales marcos de referencia para los equipos de francés en el portal PORTLEX: por un lado, la ya mencionada Teoría Sentido-Texto formulada por Mel'čuk y, por otro, la Teoría de las clases de objeto desarrollada por Gross en el *Laboratoire d'Informatique Documentaire et Linguistique* (LIDL). Ambos enfoques se han venido desarrollando, además, para impulsar avances en el ámbito del procesamiento del lenguaje natural desde hace décadas (Gross, 2004; Iordanskaja, Kim & Polguère, 1996). Resultaron de especial importancia los trabajos de Gross sobre el funcionamiento del grupo nominal en francés (Gross, 1991, 2012) y, más concretamente, sobre la noción de clase de objeto (Gross, 2002, 2008)¹⁶. Siguiendo esta línea, también se consultaron contribuciones de Blanco (1997, 1999), quien trabaja desde una perspectiva contrastiva francés-español. Finalmente, los trabajos de Lazard (1988, 1994), aunque centrados en los predicados verbales, se consultaron para evaluar el carácter actancial de algunos complementos nominales, algo para lo que también se tuvo en cuenta el estudio de Stage (1994).

Asimismo, se consultaron diferentes investigaciones más particulares sobre la estructura de los predicados, normalmente centradas en los predicados constituidos por sustantivos deverbales (Condette, Marín & Merlo, 2012; Stage, 1997). Finalmente, aunque se realizasen ya en el ámbito de la

¹⁶ En concreto, este concepto de clase de objeto, entendido como un “ensemble de substantifs, sémantiquement homogènes, qui détermine une rupture d'interprétation d'un prédicat donné, en délimitant un emploi spécifique” (Gross, 2008, p. 11), se tuvo muy presente en la creación y anotación semántica de los paquetes léxicos en los proyectos *MultiGenera* y *MultiComb*.

lingüística española, resultaron de gran utilidad los capítulos sobre el sintagma nominal de Rigau (1999) y Picallo (1999) en la *Gramática descriptiva de la lengua española*, así como la tesis de doctorado de Barrios Rodríguez (2010).

3.2. HERRAMIENTAS DE PROCESAMIENTO DEL LENGUAJE NATURAL PARA LA LENGUA FRANCESA

Además de las herramientas diseñadas específicamente para los proyectos en cuestión, el equipo de francés utilizó diferentes herramientas específicas para esta lengua disponibles en línea. Afortunadamente, para el francés hay disponibles numerosos recursos y herramientas útiles que permitieron avanzar de manera eficiente en todas las fases de trabajo. Así, contamos con corpus para la extracción de datos y la definición de prototipos léxico-semánticos, bases de datos como WordNet para desarrollar los paquetes léxicos o librerías de datos morfológicos como la facilitada en *Freeling*, esenciales para generar la flexión verbal en los contextos oracionales producidos con *CombiContext*. En diferentes casos estos recursos se usaron para alimentar herramientas de elaboración propia (Lematiza, Combina, Flexiona, etc.), también compatibles con el francés. En este apartado nos centraremos, por lo tanto, en comentar los principales recursos utilizados y las limitaciones o dificultades que presentan.

3.2.1. El trabajo con corpus y sus dificultades

Los corpus constituyeron una herramienta fundamental, tanto para la elaboración de entradas en el diccionario PORTLEX como en los proyectos subsiguientes. Permitieron obtener a gran escala datos relevantes y representaron un recurso esencial para el análisis lingüístico. En el caso del diccionario se usó FRANTEXT¹⁷ (Montémont, 2020; Pierrel, 2003) para identificar tanto argumentos nominales como sus combinaciones e ilustrar todo esto

¹⁷ <https://www.frantext.fr/>

con ejemplos. FRANTEXT es una base de datos de textos franceses¹⁸ que contiene una gran cantidad de textos literarios, científicos y técnicos desde el siglo XVIII hasta la actualidad. El corpus consta de más de 3.600 textos y 215 millones de palabras. Sin duda, es uno de los mayores corpus de textos en francés disponibles en línea y es utilizado por numerosos investigadores y estudiantes para realizar estudios lingüísticos y literarios. Este potente recurso permite localizar palabras específicas, lemas y expresiones regulares en una obra concreta o en un conjunto de fuentes.

Sin embargo, dado el carácter eminentemente literario de sus textos, en muchos casos resultaba difícil encontrar ejemplos de ciertas realizaciones argumentales. Concretamente, durante el análisis de la combinatoria argumental, FRANTEXT no permitía verificar, por falta de ejemplos, numerosas realizaciones existentes en el uso de la lengua. Por ejemplo, en el caso del lexema *consommation* ('uso de bienes o productos para la alimentación'), esta base de datos no proporcionaba ejemplos para un argumento tan habitual como el agentivo *par + determinante + nombre*, p. ej. *La consommation [de vin] par les Français*. Tampoco lo hacía para muchas combinaciones argumentales del sustantivo *apprentissage* ('acción de aprender un oficio o profesión') y ejemplos relativamente comunes como *l'apprentissage par l'enfant des structures syntaxiques* o *un apprentissage de deux ans comme consultant* tuvieron que obtenerse en corpora basados en la web.

Estas limitaciones llevaron al equipo de francés a contar exclusivamente con el corpus FrTenTen (Jakubíček, Kilgarriff, Kovář, Rychlý & Suchomel, 2013) y el interfaz de búsqueda Sketch Engine para los proyectos *MultiGenera* y *MultiComb*. Este corpus se basa en webs francesas y su versión de 2017, la utilizada en los proyectos que nos ocupan, contiene más de 10 mil millones de palabras. El uso de este potente recurso hizo posible el reconocimiento de

¹⁸ La base de datos es mantenida por el Centre National de la Recherche Scientifique (CNRS) y la Universidad de Chicago, dentro del proyecto ARTFL.

prototipos léxicos en las realizaciones argumentales mediante la extracción de datos de frecuencias. Sin embargo, el uso de corpus presenta limitaciones para obtener datos significativos sobre el comportamiento de realizaciones argumentales menos frecuentes, tanto en solitario como combinadas con otros argumentos. A este respecto cabe señalar las dificultades encontradas para obtener ejemplos válidos y suficientes para muchas realizaciones adjetivas de argumentos como, por ejemplo, en *déménagement* (*déménagement familial* = *de la famille*, ‘agente animado’) o en *mort* (*mort cancérigène* = *par cancer*, ‘causa’). Pero, sin duda, la principal limitación que presentan los corpus para el tipo de investigación que realizamos es el hecho de que ninguno está anotado semánticamente. Como veremos, esto imposibilita, entre otras cosas, la desambiguación automática de realizaciones similares de argumentos diferentes de un mismo sustantivo como, por ejemplo, *le séjour de Pierre* (‘agente’) y *le séjour de plaisance* (‘clase’).

Asimismo, surgieron dificultades en el empleo del lenguaje CQL (Corpus Query Language) (Lexical Computing, 2023a), que permite extraer datos precisos del corpus siguiendo diversos criterios. Para simplificar nuestras búsquedas y evitar obtener resultados ambiguos o carentes de relevancia, en los proyectos *MultiGenera* y *MultiComb* se optó por extraer únicamente datos relacionados con la primera posición argumental de los predicados nominales, es decir, la posición contigua a su núcleo¹⁹. La inclusión de otras posiciones de los argumentos en nuestra investigación habría requerido la definición de expresiones regulares mucho más complejas y el procesamiento de un volumen de datos difícil de asumir. La Tabla 1 muestra las estructuras para las que se extrajeron datos.

¹⁹ Cabe recordar aquí que en estos proyectos la finalidad de nuestras búsquedas en corpus no era un estudio exhaustivo de los predicados nominales de los sustantivos escogidos, sino la identificación de prototipos léxicos en cada argumento nominal para proceder a la elaboración semiautomática de paquetes semánticos (Valcárcel Riveiro & Pino Serrano, 2023).

Estructuras consultadas	Ejemplos
Det. + Núcleo + Prep. + Nombre	<i>Le voyage de Marie</i> <i>Le voyage en Italie</i>
Det. + Núcleo + Prep. + Det. + Nombre	<i>Le voyage de la professeure</i> <i>Le voyage depuis le Japon</i>
Det. + Núcleo + Adjetivo	<i>Le voyage présidentiel</i> <i>Le voyage asiatique</i>

TABLA 1: Estructuras consultadas en Sketch Engine para los proyectos MultiGenera y MultiComb

Esta decisión de analizar sólo las estructuras contiguas al núcleo implica que los KWIC²⁰ extraídos para el análisis presenten problemas de interpretación semántica. La frecuencia de estos casos hace inviable la consulta directa de los ejemplos en el corpus para su desambiguación, por lo que se adoptaron varias soluciones de compromiso. Así, los KWIC no muestran los núcleos argumentales cuando en esta posición se encuentran adjetivos o sustantivos compuestos sin guiones. Por ejemplo, cuando en el núcleo del argumento aparecen *jeune fille* o *petit ami* en el KWIC solo vemos *la présence de la jeune*, *la vidéo de son petit*. En estos casos se analizan los elementos que aparecen en el KWIC (*jeune*, *petit*) como sustantivos. Cuando resulta obvio que se trata de un adjetivo (p.ej. *la conversation avec ce beau [garçon]*), se descarta el KWIC. Un problema similar también hizo inviable el procesamiento de sintagmas nominales con determinantes cuantificadores adverbiales (p.ej., *beaucoup de*, *peu de*) o nominales (*la moitié de*, *un tas de*, *la majorité de*) (Gross, 2012, pp. 177-178)²¹ ya que los KWIC no mostraban el

²⁰ KWIC es un acrónimo en inglés que significa *Key Word In Context*. En Sketch Engine, un KWIC se obtiene mediante una expresión regular en CQL y permite visualizar la palabra o expresión buscada en un contexto más amplio. Esto facilita el análisis contextual de los resultados de las búsquedas en el corpus. La palabra clave se muestra en el centro de la pantalla, rodeada de las palabras que la acompañan en las oraciones donde aparece (Lexical Computing, 2023b).

²¹ Wilmet (1997, pp. 168-171, 227-229) analizó en detalle estas estructuras clasificándolos como cuantificadores estrictos compuestos (p. ej. *un morceau de*, *beaucoup de*, *assez de*) o como cuantificadores-caracterizantes preposicionales (p. ej. *une sorte de*, *une espèce de*).

núcleo del argumento, p.ej.: *la présence de la majorité, la largeur de beaucoup de, l'odeur de la plupart*.

Asimismo, esta limitación de contexto en las búsquedas dificultó la desambiguación de palabras polisémicas. Por ejemplo, en el KWIC *le voyage de la femme* es imposible saber, sin verificarlo en el ejemplo fuente, si aquí el sustantivo *femme* es sinónimo de *épouse* (*la femme de Carla*) o si se refiere al género de una persona (*la femme de l'épicerie*). Sin embargo, en KWIC *le voyage de sa femme* el posesivo indica claramente que el significado es 'esposa'. Para zanjar el problema, estos casos ambiguos se anotaron semánticamente con la etiqueta más general: en el caso de *femme* como 'ser humano femenino'.

Las dificultades en el análisis semántico pueden aparecer también en algunos núcleos nominales como *présence* o *mort*. En estos casos se tuvieron que descartar las búsquedas con determinantes en plural (por ejemplo, *les présences, les morts*) porque esto podría interpretarse de diferentes maneras. Así, en el caso de *présence*, la forma plural *les présences* indica más bien 'persona o entidad presente' (p.ej., *les présences dans la salle de réunion étaient nombreuses*) o más raramente 'momento en el que alguien está presente' (p.ej., *ses présences dans la salle de classe étaient toujours appréciées par les élèves*), y no ya el significado estudiado en los proyectos referente al 'hecho de encontrarse en un lugar'.

3.2.2. WordNet y las herramientas PORTLEX

Los proyectos MultiGenera y MultiComb demandaban la extracción de una gran cantidad de datos léxicos etiquetados semánticamente. Con ellos se elaboraron paquetes semánticos para alimentar los prototipos que generaban frases nominales personalizadas de manera automática (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019; Domínguez Vázquez, Bardanca Outeiriño & Simões, 2021). WordNet respondía en gran medida a las necesidades de los proyectos y resultó esencial para su desarrollo desde un primer momento. Esto se debe, sin duda, al particular tratamiento de la

información semántica en WordNet. Esta sólida base de datos léxica organiza semánticamente palabras en conjuntos de sinónimos conocidos como *synsets* (Miller, Beckwith, Fellbaum, Gross & Miller, 1990). Sin embargo, este recurso lingüístico va más allá de un mero diccionario, ya que facilita el análisis automático de textos desde un punto de vista semántico y, por tanto, el desarrollo de herramientas como las que se crearon en los proyectos *Multi-Genera* y *MultiComb*.

Otro factor esencial es la amplia cobertura multilingüe de WordNet, lo que garantiza la disponibilidad de una gran cantidad de datos léxicos anotados semánticamente para todas las lenguas de trabajo en los proyectos desarrollados en PORTLEX: castellano, gallego, francés, italiano y alemán. Más concretamente, en el caso del francés se utilizó la versión 1.0b4 del *WordNet Libre du Français* (WOLF) (Sagot & Fišer, 2008)²². Esencialmente, WOLF se construyó a partir del WordNet original desarrollado en Princeton. A lo largo del tiempo, se ha ido ampliando y mejorando en diferentes versiones en las que se han procesado polisemias y nominalizaciones, recurriendo incluso a la depuración manual de datos inconsistentes. Este recurso se encuentra en formato XML y se puede descargar libremente en la web del proyecto²³.

Sin embargo, aunque la cobertura léxica de esta herramienta respondía a las necesidades de los proyectos *MultiGenera* y *MultiComb*, la información semántica asociada en el WOLF a las unidades de significado o *synsets* es limitada. Esta carencia se suplió mediante un emparejamiento con los datos de EuroWordNet²⁴ del Multilingual Central Repository (MCR), ya que

²² El WOLF fue desarrollado en Francia por el equipo ALMAAnaCH, que trabaja en el seno del Institut National de Recherche en Informatique et en Automatique (INRIA).

²³ https://almanach.inria.fr/software_and_resources/WOLF-fr.html

²⁴ EuroWordNet es una base de datos multilingüe que incluye WordNets en varios idiomas europeos. Cada idioma estructura su propio WordNet en *synsets* y relaciones semánticas básicas entre ellos siguiendo una organización similar al WordNet original de Princeton.

estos están asociados a información semántica categorizada en diferentes ontologías como Suggested Upper Merged Ontology (SUMO), Top Concept Ontology (Top), WordNet Domains, Basic Level Concept y Epinónimos. Este paso tan importante se realizó en colaboración con el equipo de GalNet²⁵, la versión de WordNet para el gallego (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019; Gómez Guinovart & Solla Portela, 2020). Gracias a esto, resultó posible hacer búsquedas semánticas de datos léxicos y refinarlas, además, por tipos de palabra. De esta manera se extrajeron de manera masiva los datos necesarios para crear paquetes anotados semánticamente siguiendo un sistema de etiquetado concebido *ad hoc* para los proyectos del portal PORTLEX (Domínguez Vázquez, Valcárcel Riveiro & Bardanca Outeiriño, 2021). Sin embargo, para poder realizar la extracción de datos en Galnet de manera efectiva se desarrollaron diferentes herramientas específicas.

En primer lugar, se diseñaron diferentes APIs de consulta a Galnet, siendo una de ellas específica para el francés. Su función es la de obtener datos léxicos filtrados por categorías semánticas de las diferentes ontologías (SUMO, TOP, etc.). Esta API proporciona resultados con gran rapidez en formato JSON, lo que facilita la utilización de los datos obtenidos en otros recursos. Sin embargo, los datos proporcionados por la API se limitaban a una única categoría de una ontología concreta. Esto suponía un problema, ya que estas categorías ontológicas estaban organizadas por criterios semánticos generales y, por lo tanto, no atendían al contexto de uso de cada palabra o lexema, esto es, a su predicado y a las combinatorias que se dan en él. Así,

Estos WordNets están conectados a través de un índice interlingüe (ILI), lo que permite consultar lexemas similares en otros idiomas y acceder a una ontología compartida con 63 distinciones semánticas comunes a todos los idiomas. El proyecto EuroWordNet comenzó en 1994 y se completó en 1999, estableciendo una base sólida para la expansión de recursos lingüísticos en diferentes idiomas europeos (Vossen, 1999).

²⁵ GalNet es un recurso desarrollado por el Seminario de Lingüística Informática de la Universidad de Vigo. Este proyecto se enmarca en un esfuerzo más amplio de integrar de forma coordinada las versiones del WordNet 3.0 en español, catalán, gallego, vasco y portugués. (Gómez Guinovart & Solla Portela, 2018, 2020).

dos substantivos como *pain* (07679356-n)²⁶ y *farine* (07567707-n)²⁷ aparecen incluidos en la misma categoría *Food* de la ontología SUMO, también en *Comestible* de TOP o en *Gastronomy* de WordNet Domains. Sin embargo, desde un punto de vista sintáctico-semántico *pain* y *farine* pertenecen a clases o categorías diferentes puesto que sus rasgos semánticos no permiten la combinación con verbos o adjetivos similares²⁸. En términos de Gross (2008, 2012), *pain* y *farine* pertenecen a dos clases de objetos diferentes y, aunque ambos lexemas comparten rasgos semánticos como +material +comestible, *pain* es una comida y *farine* un ingrediente. Así, en nuestros prototipos de *Combinatoria* y *Xera* (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019; Domínguez Vázquez, Bardanca Outeiriño & Simões, 2021) dejar estos dos substantivos en un mismo paquete semántico llevaría a la generación de frases nominales semánticamente inaceptables como **le goût de la farine* o **la largeur de la farine*. Para evitar este tipo de resultados y mantener la automatización del proceso de importación de datos se necesitó más granularidad semántica en su extracción desde WordNet.

Para dar respuesta a esta necesidad, se desarrolló la herramienta *Combina*, que permite combinar múltiples consultas semánticas en Galnet para el mismo idioma, ya sea mediante la adición de datos de una consulta inicial a otra o la intersección de resultados de diferentes consultas (ver Figura 1). Los datos resultantes se generan tanto en formato de texto como en JSON. Por ejemplo, con esta herramienta podemos obtener los sustantivos compartidos en las categorías *Gastronomy* de WordNet Domains, *Food* de SUMO y *Artifact* de TOP con el fin de elaborar un paquete semántico de platos o

²⁶ La glosa o definición proporcionada en WordNet para este *synset* es “food made from dough of flour or meal and usually raised with yeast or baking powder and then baked”.

²⁷ La glosa en inglés para este *synset* es “coarsely ground foodstuff; especially seeds of various cereal grasses or pulse”.

²⁸ Así, por ejemplo, *pain* no puede ser objeto directo de verbos como *saupoudrer* o *tamiser*, ni *farine* puede ser objeto directo de los verbos *couper*, *griller* o *manger*. Por otro lado, *farine* no puede adjetivarse con *rassise* o *croustillante*, ni *pain* con *moulu*.

productos gastronómicos en el que se encuentre *pain*, pero no *farine*. Esta búsqueda cruzada en *Combina* nos proporciona 411 lemas entre los que encontramos *bouillon*, *confiture*, *fondue*, *gâteau*, *quiche*, *pâté* o *ragoût*, pero no *farine*, *oeuf* o *cannelle*²⁹.

The screenshot shows the Combina web interface with the following elements:

- Logos:** MultiGenera, MultiComb, USC (Universidad del Sur de California), Fundación BBVA, and various institutional logos.
- Language Selection:** A section titled "Seleccione a lingua de traballo" with radio buttons for "Deutsch", "español", and "français" (which is selected).
- Combination Type:** A section titled "Tipo de combinación" with radio buttons for "lemas combinados" and "lemas compartidos" (which is selected).
- Results:** A section titled "Resultados" with radio buttons for "Todos", "substantivos" (selected), "adjectivos", "verbos", and "adverbios".
- APIs:** Four API endpoints are listed:
 - API 1: <http://portlex.usc.gal/develop/fr/api/?ontology=domains&category=gastronomy>
 - API 2: <http://portlex.usc.gal/develop/fr/api/?ontology=sumo&category=Food>
 - API 3: <http://portlex.usc.gal/develop/fr/api/?ontology=top&category=Artifact>
 - API 4: <http://portlex.usc.gal/develop/fr/api/>

FIGURA 1: Vista de la interfaz de Combina en una búsqueda de datos compartidos por tres categorías ontológicas en WordNet

Naturalmente, extraer datos precisos con Combina requiere tener un buen conocimiento de la estructura de las diferentes ontologías asociadas a WordNet. Dado que este no era el caso de los integrantes de los equipos de lingüistas, entre ellos el de francés, se requirió el desarrollo de una nueva herramienta que facilitase rápidamente la localización de un determinado lexema en las categorías de las diferentes ontologías. Esta nueva herramienta se denominó Lematiza y acepta concordancias de los corpus disponibles en

²⁹ De todas formas, aunque con *Combina* se obtienen datos mucho más precisos, su revisión y depuración humana es todavía necesaria. Por ejemplo, entre los resultados obtenidos en la búsqueda mencionada antes se incluyen ítems repetidos como *purée*, que se asocian a varios *synsets* o significados en WordNet, o sustantivos como *assiette* ou *viennoiserie*, que no responden completamente a la delimitación semántica del paquete que pretendemos crear.

Sketch Engine o listas de frecuencias en formatos csv y xml (ver Figura 2). Al ejecutarla, se obtienen los lemas de estas consultas, con sus diferentes *synsets* y las categorías ontológicas a las que se asocian en WordNet. Además, la herramienta proporciona enlaces de la API correspondientes a consultas de ontologías externas relacionadas, así como a otros datos de tipo semántico (hipónimo, hiperónimos, etc.) (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019).

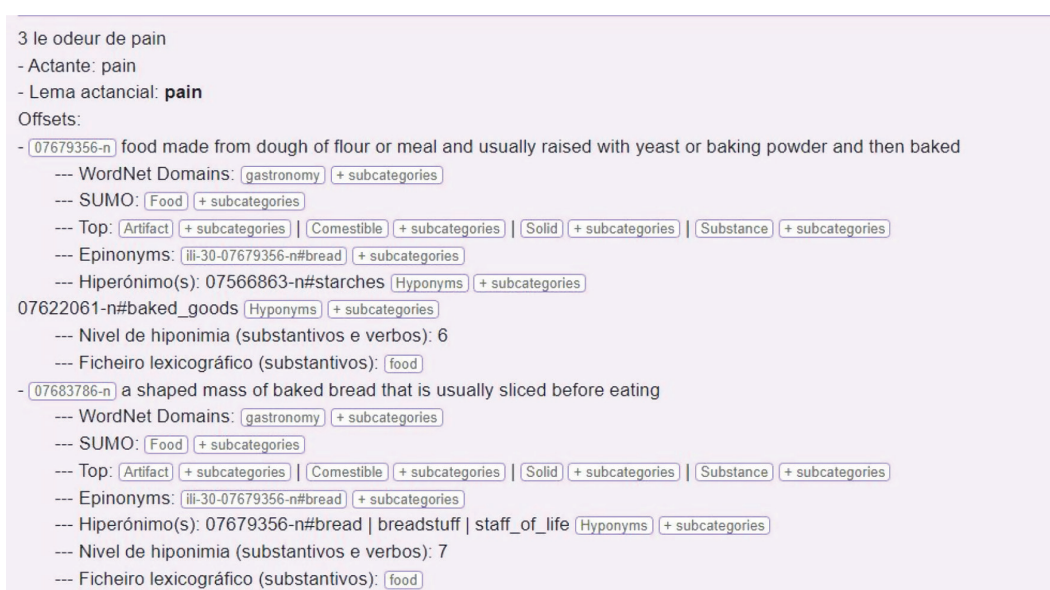


FIGURA 2: Vista de los resultados proporcionados por Lemmatiza para los tres synsets del lema pain

3.2.3. FreeLing y el tratamiento de los paradigmas morfológicos

Además de la extracción masiva de datos léxicos y su organización semántica, tanto *Xera* como los prototipos de combinatoria frasal y verbal requerían información morfológica detallada para generar tanto sintagmas nominales como sus contextos oracionales. En el prototipo Combinatoria se necesitaba flexionar el género y número de sustantivos y las correspondientes concordancias en determinantes y adjetivos, así como conjugar los verbos en *CombiContext*. Entre las diferentes opciones disponibles para cada idioma, se priorizó la búsqueda de un recurso multilingüe. De esta manera, se optó

finalmente por FreeLing, una biblioteca desarrollada en C++ que proporciona un conjunto de recursos multilingües para el procesamiento del lenguaje natural. Entre sus diferentes funcionalidades se encuentra el análisis morfológico de numerosas lenguas, incluido el francés. De esta forma FreeLing proporciona información morfológica detallada sobre la flexión de sustantivos, adjetivos, determinantes y verbos (Padró & Stanilovsky, 2012).

De la biblioteca de FreeLing no se extrajeron todos los datos disponibles para el francés y las otras lenguas de trabajo, sino solamente las formas flexionadas de los sustantivos y adjetivos presentes en los diferentes paquetes semánticos. Para automatizar la extracción de la información morfológica de FreeLing y su incorporación en los paquetes semánticos obtenidos con *Combina* se diseñó otra herramienta: Flexiona (Domínguez Vázquez, Solla Portela & Valcárcel Riveiro, 2019). Gracias a ella, seleccionando la lengua de consulta y cargando el paquete semántico correspondiente, se realizaba rápidamente esta operación. En los paquetes, las formas flexionadas importadas se asocian al lema y a su correspondiente *synset*. Se indica primero la forma flexionada, después el lema, los códigos de tipo de palabra, género y número, y finalmente el *synset*. Los ejemplos (1) y (2) muestran la disposición de los datos en columnas correspondientes al sustantivo *maladie* y al adjetivo *violent*.

(1)	maladie	maladie	N	F	S	14061805-n
	maladies	maladie	N	F	P	14061805-n
(2)	violent	violent	A	M	S	02510879-a
	violents	violent	A	M	P	02510879-a
	violente	violent	A	F	S	02510879-a
	violentes	violent	A	F	P	02510879-a

Aunque esta automatización de la anotación morfológica de los paquetes léxicos fue satisfactoria para la mayoría de los lemas, surgieron problemas en diferentes casos:

- a. Los lemas multipalabra (p. ej. *peste noire*, *fruit de la passion*) no se incluyen en la biblioteca de Freeling y, por lo tanto, no se pudo importar para ellos la información morfológica. Esta se tuvo que introducir manualmente durante la elaboración de los paquetes semánticos.
- b. Las formas femeninas de singular y plural tampoco constaban en Freeling para muchas profesiones (p. ej. *écrivain*, *professeur*, *transporteur*), por lo que también hubo que incorporarlas de forma manual a los paquetes de datos. Para verificar las formas femeninas de estos sustantivos (p. ej., *écrivaine*, *professeure*, *transporteuse*) se consultó la guía *Femme, j'écris ton nom*, publicado por el CNRS y el Institut de la Langue Française (Cerquiglini, 1999).
- c. En los argumentos locativos (p. ej. *le voyage en Italie*, *le séjour au Portugal*) surge el problema de la variación de la preposición *en* ~ *à*, vinculada generalmente al tipo de lugar y, en el caso de los países, a su género. Aunque Freeling proporcionaba los datos de género para los países, hubo que crear paquetes específicos³⁰ ya que en realidad la regla contempla no pocas excepciones. Así, entre los países que rigen *en* podemos encontrar nombres masculinos (p. ej., *Israël*, *Iran*) y entre los países que rigen *à* + determinante nombres femeninos (p. ej., *Philippines*, *Seychelles*). Esto llevó, por un lado, a la revisión manual de los paquetes y a la programación de estructuras diferentes para cada preposición, es decir, procesarlas como realizaciones diferentes del mismo argumento.
- d. Las palabras con la llamada “h aspirada” no se identifican en las bibliotecas de Freeling y esto complicó la programación de los prototipos. Hubo que crear código específico para evitar que en las frases con

³⁰ Se elaboraron separadamente tres paquetes de nombres de países que regían la preposición *à* (p. ej. *Cuba*, *Chypre*), nombres de países que regían *à* + determinante (p. ej. *Yémen*, *Portugal*) y países o regiones que regían la preposición *en* (p. ej. *Italie*, *Bretagne*).

palabras con h aspirado no se generasen elisiones ni se colocasen apóstrofes (p. ej. *le hamburger* y no **l'hamburger*, *la Hongrie* y no **L'Hongrie*).

Como se puede ver, todas estas cuestiones hicieron necesaria una revisión manual y exhaustiva de todos los paquetes semánticos básicos obtenidos de *Combina* (92), lo que implicó una importante inversión de tiempo por parte del equipo de trabajo³¹. La necesidad de concentrar tiempo y esfuerzos en la revisión de los paquetes llevó a limitar la morfología de los determinantes y de los verbos con el fin de evitar más complicaciones en la generación automática. En el caso de los determinantes, se decidió limitar al artículo determinado la generación de estos en los sintagmas nominales. Esto se debe a las numerosas restricciones a las que está sujeto el uso de determinantes en estas estructuras (Gross, 1991, pp. 269-270). Esta decisión ahorró eventuales revisiones de las frases generadas, pero llevó a generar grupos nominales poco frecuentes en el uso, véase:

(3) *Les voyages de la mère sont fréquents.*

(4) *La largeur de la tête est surprenante.*

En los dos ejemplos *mère* y *tête* son dos sustantivos asociados a la propiedad inalienable de alguien, que se expresa mediante un complemento preposicional o un determinante posesivo (Rigau, 1999, pp. 345-346). La decisión de limitar la generación de determinantes a los artículos definidos excluía obtener en estos casos resultados más comunes:

(5) *Les voyages de ta mère sont fréquents.*

(6) *La largeur de sa tête est surprenante.*

³¹ Cabe recordar aquí que la elaboración de los paquetes conllevó dos revisiones manuales además de esta: una anterior, en la que se depuraban los lexemas no pertinentes extraídos de WordNet, y otra posterior en la que se depuraban los lexemas inapropiados para la combinatoria con un determinado sustantivo. Así, el paquete *fr_animado_planta_arbol* que se combina con el argumento 'tipo' del sustantivo *odeur* no contiene lexemas como *saule* o *bêtre* (p. ej., **l'odeur de saule*) pero sí el que se combina con el argumento 'tema' de *conversation* (p. ej. *la conversation sur le saule*).

En el caso de los verbos, se limitó la generación de la conjugación a las terceras personas del presente del indicativo. Se excluyó, por un lado, el uso de pronombres personales sujeto (*je, tu, elle*, etc.), y por el otro la generación de verbos en pasado, en futuro o en imperativo. Así, el número de formas importadas de Freeling para cada verbo se limita a tres (infinitivo, tercera persona del singular y tercera persona del plural), p. ej. *penser, pense, pensent*. Esto permitió reducir considerablemente el número de datos almacenados en la base de datos y simplificar eventuales revisiones.

4. CONCLUSIONES

En la presente contribución se ha hecho un repaso al trabajo desarrollado sobre el francés en el portal lexicográfico PORTLEX. Los equipos que trabajaron con esta lengua en los tres proyectos realizados (el diccionario PORTLEX, *MultiGenera* y *MultiComb*) se enfrentaron a importantes desafíos relacionados, sobre todo, con la obtención y procesamiento de grandes cantidades de datos léxicos. Para ello se usaron una serie de herramientas y recursos procedentes de diferentes ámbitos como la lingüística computacional (corpus, bases de datos, repositorios), la lexicografía (diccionarios, léxicos) o la gramática (gramáticas, artículos de investigación). Además, dentro del portal se desarrollaron nuevas aplicaciones para simplificar el trabajo de los equipos de lingüistas.

Sin embargo, a pesar de todas las herramientas disponibles, los proyectos PORTLEX requirieron mucho tiempo de trabajo por parte de estos equipos. La particular naturaleza valencial y multilingüe del diccionario y los prototipos de generación desarrollados exigió un minucioso trabajo previo de análisis lingüístico y depuración de datos léxicos imposible de automatizar completamente. A esto se le sumaron las diferentes limitaciones que presentan las herramientas y recursos empleados y que se detallan en la parte central del artículo. Afortunadamente, todos estos problemas no supusieron una interrupción del trabajo ya que se fueron encontrando soluciones que los

resolvían o los paliaban. De esta forma, se consiguieron alcanzar los principales objetivos en cada proyecto y, finalmente, tanto el diccionario PORTLEX como los prototipos de generación *Combinatoria* y *CombiContext* son una realidad. De hecho, los resultados alcanzados para el francés se pueden consultar en acceso libre y de manera contrastada con los datos de otras lenguas de trabajo.

Todo lo aprendido en estos tres proyectos redundará, sin duda, en una mayor eficacia de los equipos a la hora de afrontar el nuevo desafío que supone ESMAS-ES+. De hecho, este proyecto, además de presentar la orientación multilingüe y valencial que caracteriza el trabajo en el portal PORTLEX desde sus inicios, se define por una vocación de sostenibilidad. Esto implicará la reutilización de datos y herramientas de los proyectos anteriores para desarrollar un etiquetador semántico evitando el desperdicio de tiempo, material y dinero. En última instancia, se pretende que estos esfuerzos contribuyan a impulsar un futuro más eficiente y sostenible en la investigación lingüística y el procesamiento del lenguaje natural.

REFERENCIAS BIBLIOGRÁFICAS

- Abeillé, A. & Godard, D. (dirs.) (2021). *La grande grammaire du français*. Actes Sud.
- Alonso-Ramos, M. (2004). *Diccionario de colocaciones del español (DICE)*. <http://www.dicesp.com/paginas>
- Alonso-Ramos, M. (2017). Diccionarios combinatorios. *Estudios de Lingüística del Español*, 38, 173-201. <https://doi.org/10.36950/elies.2017.38.8651>
- Barrios Rodríguez, M. A. (2010). El dominio de las funciones léxicas en el marco de la Teoría Sentido-Texto. *Estudios de Lingüística del español (ELiEs)*, 30. <http://elies.rediris.es/elies30/index30.html>
- Blanco, X. (1997). De las clases de objetos a las clases de predicados. *Verba*, 24, 371-385.
- Blanco, X. (1999). *Lexicographie bilingue français-espagnol et classes d'objets*. Universitat Autònoma de Barcelona.
- Cerquiglini, B. (dir.) (1999). *Femme, j'écris ton nom... Guide d'aide à la féminisation des noms de métiers, titres, grades et fonctions*. Centre National de la Recherche Scientifique & Institut de la Langue Française. https://medias.vie-publique.fr/data_storage/s3/rapport/pdf/994001174.pdf

- Charaudeau, P. (1992). *Grammaire du sens et de l'expression*. Hachette-Éducation.
- Condette, M. H., Marín, R. & Merlo, A. (2012). La structure argumentale des noms déverbaux: du corpus au lexique et du lexique au corpus. En F. Neveu, V. M. Toke, P. Blumenthal, T. Klingler, P. Ligas, S. Prévost & S. Teston-Bonnard (eds.), *Actes du 3ème Congrès Mondial de Linguistique Française, Lyon, France, 4-7 juillet 2012* (pp. 845-858). SHS Web of conferences. <https://doi.org/10.1051/shsconf/20120100271>.
- Dendien, J. (2020). *Le TLFi Trésor de la langue française informatisé*. Analyse et traitement informatique de la langue française. <http://atilf.atilf.fr/tlfr3.htm>
- Domínguez Vázquez, M.^a J. (2011). *Kontrastive Grammatik und Lexikographie: spanisch-deutsches Wörterbuch zur Valenz des Nomens*. Iudicium
- Domínguez Vázquez, M.^a J. (2018). Was sind Valenzwörterbücher? *Sprachwissenschaft*, 43(3), 309-342.
- Domínguez Vázquez, M.^a J., Bardanca Outeiriño, D. & Simões, A. (2021). Automatic Lexicographic Content Creation: Automating Multilingual Resources Development for Lexicographers. En I. Kosem, M. Cukr, M. Jakubiček, J. Kallas, S. Krek & C. Tiberius (eds.), *Post-editing Lexicography – Elex 2021. Proceedings of the eLex 2021 conference* (pp. 269-287). European Lexicographic Infrastructure. https://elex.link/elex2021/wp-content/uploads/eLex_2021-proceedings.pdf
- Domínguez Vázquez, M.^a J., Solla Portela, M. A. & Valcárcel Riveiro, C. (2019). Resources interoperability: exploiting lexicographic data to automatically generate dictionary examples. En I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubiček, S. Krek, C. Tiberius (eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal* (pp. 51-71). Lexical Computing CZ, s.r.o.
- Domínguez Vázquez, M.^a J. & Valcárcel Riveiro, C. (2020). PORTLEX as a multilingual and cross-lingual online dictionary. En M.^a J. Domínguez Vázquez, M. Mirazo Balsa & C. Valcárcel Riveiro (eds.), *Studies on multilingual lexicography* (pp. 135-158). De Gruyter. <https://doi.org/10.1515/9783110607659-008>
- Domínguez Vázquez, M.^a J., Valcárcel Riveiro, C. & Bardanca Outeiriño, D. (2021), *Ontología léxica*. Santiago de Compostela. <http://portlex.usc.gal/ontologia>
- Domínguez Vázquez, M.^a J., Valcárcel Riveiro, C. & Lindemann, D. (2018). Multilingual Generation of Noun Valency Patterns for Extracting Syntactic-Semantical Knowledge from Corpora (*MultiGenera*). En J. Cibej, V. Gorjanc, I. Kosem & S. Krek (eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana, Slovenia* (pp. 847-854). Ljubljana University Press.
- Engel, U. (2004). *Deutsche Grammatik – Neubearbeitung*. Iudicium.
- Gómez Guinovart, X. & Solla Portela, M. A. (2018). Building the Galician WordNet: methods and applications. *Language Resources & Evaluation* 52(1), 317-339. <https://doi.org/10.1515/9783110607659-010>
- Gómez Guinovart, X. & Solla Portela, M. A. (2020). Construction of a WordNet-based multilingual lexical ontology for Galician. En M.^a J. Domínguez Vázquez, M. Mirazo Balsa

- & C. Valcárcel Riveiro (eds.), *Studies on multilingual lexicography* (pp. 179-196). De Gruyter.
- Grevisse, M. & Goosse, A. (2008). *Le bon usage. Grammaire française*. De Boeck-Duculot.
- Gross, G. (1991). Syntaxe du complément de nom. *Linguisticae Investigationes*, 15, 255-284. <https://doi.org/10.1075/li.15.2.02gro>
- Gross, G. (2002). Analyse de compléments du nom en termes de classes d'objets. *Le français moderne*, 70(2), 187-209.
- Gross, G. (2004). Réflexions sur le traitement automatique des langues. En G. Purnelle, C. Fairon & A. Dister (eds.), *Le Poids des mots. Actes des 7es Journées internationales d'Analyse Statistique des Données Textuelles (JADT 2004)* (545-556). Presses Universitaires de Louvain.
- Gross, G. (2008). Les classes d'objets. *Lalies*, 28, 111-165.
- Gross, G. (2012). *Manuel d'analyse linguistique*. Presses Universitaires du Septentrion. <https://doi.org/10.4000/books.septentrion.115128>
- Iordanskaja, L., Kim, M. & Polguère, A. (1996). Some Procedural Problems in the Implementation of Lexical Functions for Text Generation. En L. Wanner (ed.), *Lexical functions in lexicography and natural language processing* (pp. 279-297). John Benjamins.
- Jakubiček, M., Kilgariff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The Tenten Corpus Family. En A. Hardie & R. Love (eds.), *Proceedings of the 7th International Corpus Linguistics Conference CL* (pp. 125127). Lancaster University. https://www.sketchengine.eu/wp-content/uploads/The_TenTen_Corpus_2013.pdf
- Lazard, G. (1988). Définition des actants dans les langues européennes. En J. Feuillet (ed.), *Actance et valence dans les langues de l'Europe* (pp. 11-146). Mouton de Gruyter. <https://doi.org/10.1515/9783110804485.11>
- Lazard, G. (1994). *L'actance*. Presses Universitaires de France.
- Lexical Computing (2023a). CQL - Corpus Query Language. *Sketch Engine*. <https://www.sketchengine.eu/documentation/corpus-querying/>
- Lexical Computing (2023b). Concordance - a tool to search corpus. *Sketch Engine*. <https://www.sketchengine.eu/guide/concordance-a-tool-to-search-a-corpus/>
- Mel'čuk, I. (1997). *Vers une linguistique Sens-Texte. Leçon inaugurale*. Collège de France. <http://olst.ling.umontreal.ca/pdf/melcukColldeFr.pdf>
- Mel'čuk, I., Arbatchewsky-Jumarie, N., Iordanskaja, L., Mantha, S. & Polguère, A. (eds.) (1984-1999). *Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques I-IV*. Presses de l'Université de Montréal. <https://doi.org/10.2307/j.ctv69t5n2>
- Mel'čuk, I., Clas, A. & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Éditions Duculot.
- Mel'čuk, I. & Milićević, J. (2014). *Introduction à la linguistique. Volume 1*. Éditions Hermann.

- Mel'čuk, I. & Polguère, A. (2007). *Lexique actif du français: l'apprentissage du vocabulaire fondé sur 20000 dérivations sémantiques et collocations du français*. De Boeck.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D & Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4), 235-244. <https://doi.org/10.1093/ijl/3.4.235>
- Montémont, V. (2020). De Frantext 1 à Frantext 2: la cure de jouvence d'une vieille dame. En D. Aquino-Weber & Y. Greub (eds.), *La lexicographie informatisée: les vocabulaires nationaux dans un contexte européen* (pp. 41-66). Académie suisse des sciences humaines et sociales.
- Padró, L. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. En N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. European Language Resources Association. <https://nlp.lsi.upc.edu/publications/papers/padro12.pdf>
- Picallo, M. C. (1999). La estructura del sintagma nominal: las nominalizaciones y otros sustantivos con complementos argumentales. En I. Bosque & V. Demonte (dirs.), *Gramática descriptiva de la lengua española* (pp. 363-393). Espasa Calpe.
- Pierrel, J. M. (2003). Un ensemble de ressources de référence pour l'étude du français: TLFI, FRANTEXT et le logiciel STELLA. *Revue québécoise de linguistique*, 32, 155-176. <https://doi.org/10.7202/012248ar>
- Riegel M., Pellat, J.C. & Rioul, R. (2009). *Grammaire méthodique du français*. Presses Universitaires de France.
- Rigau, G. (1999). La estructura del sintagma nominal: los modificadores del nombre. En I. Bosque & V. Demonte (dirs.), *Gramática descriptiva de la lengua española* (pp. 311-362). Espasa Calpe.
- Sagot, B. & Fišer, D. (2008). Building a free French wordnet from multilingual resources. En N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (eds.), *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. European Language Resources Association. <https://inria.hal.science/inria-00614708/document>
- Stage, L. (1994). La valence des noms en français, En M. Herslund (ed.), *Noun Phrase Structures* (pp. 93-131). Samfundslitteratur.
- Stage, L. (1997). La transposition des actants dans le syntagme nominal. Étude sur la nominalisation nucléaire et l'emploi des prépositions. *Revue Romane*, 32(1), 51-86.
- Tesnière, L. (1959). *Éléments de syntaxe structurale*. Klincksieck.
- Valcárcel Riveiro, C. (2017). Las construcciones N_1N_2 como realizaciones actanciales del sustantivo en francés y su tratamiento en el diccionario multilingüe PORT-LEX. En M.^a J. Domínguez Vázquez & S. Kutscher (eds.), *Interacción entre gramática, didáctica y lexicografía* (pp. 193-207). De Gruyter. <https://doi.org/10.1515/9783110420784-015>

- Valcárcel Riveiro, C. & Pino Serrano, L. (2023). Application d'une méthodologie d'analyse des prédicats nominaux: l'exemple du lexème MORT₁. *Çédille. Revista de estudios franceses*, 24 (en prensa).
- Vossen, P. (ed.) (1999). *EuroWordNet. General Document 1, Final*. University of Amsterdam. <https://archive.illc.uva.nl/EuroWordNet/docs/GeneralDocDOC.zip>
- Wilmet, M. (1997). *Grammaire critique du français*. Duculot.