Workflow and problems of a multilingual electronic dictionary: the case of PORTLEX

María José Domínguez Vázquez, Mónica Mirazo Balsa, Carlos Valcárcel Riveiro
Universidad de Santiago de Compostela, Universidad de Santiago de Compostela, Universidad de Vigo
E-mail: majo.dominguez@usc.es, monica.mirazo@usc.es, carlos.valcarcel@uvigo.es

1. Introduction

The aim of our contribution is to describe the workflow of a research project entitled: *Lexicographic Portal: Modular Online Multilingual Dictionary Based On An Annotated Corpus Of The Noun Phrase*.[1] The main goal of this project is the making of an annotated, computerized Spanish – German – Galician – Italian - French corpus and its management systems for the analysis of the noun phrase and its combinations using valency parameters, as well as its implementation in a multilingual lexicographic online dictionary. This is a dictionary under construction and this research is based on the valence theory, on the analysis of vast textual corpora (which provides a solid empirical basis for the lexicographic description) and on previous results of the CSVEA project (INCITE09204074 PR). This new project has an interdisciplinary nature with contributions, not only from different language domains, but also from translation studies as well as from computational and corpus linguistics, providing an important number of analytic combinations and future applications (monolingual, contrastive and interlingual aspects).

We are thus speaking of an innovative lexicographic portal with plenty of search and information options, based on three central issues: a) multiplicity of languages and a detailed reversible contrastive valential information, b) attention to the online architecture of the information, to the diverse user typologies, as well as to the new analytical methods in lexicography and c) the modular nature of the web portal.

The goals set are: 1. The making of an annotated multilingual corpus and a modular, reversible and multilingual lexicographic online portal, based on valency grammar and lexicography, 2. The development of monolingual and reversible contrastive dictionaries for the different languages concerned, based on the exploitation of textual corpora as an empirical base for the lexicographic work and on the research on user typology and information architecture, 3. The inclusion of other languages and word classes.

Among the main problems posed by this project, we can highlight the following:

- The design of an adequate database for the multilingual nature of the project, particularly in terms of data management and retrieval.
- The coordination of researchers' work on different language and lexicographical traditions.
- The collection and management of a high amount of data on a relatively short schedule.

2. Lexicographical workflow

2.1. Preparation

We could differentiate two preparation phases:

Preparation phase A: As the research is done in a university, it is necessary to seek approval of the project at a public institution (ministry, regional government), on which the official status of the project and its funding are heavily reliant. Projects are granted typically for a maximum period of three years, and this affects the planning. This stage involves tasks such as the project conception, the organizational plan, several meetings or the project application.

Preparation phase B: At this stage a proposal concerning the descriptive model and the metalanguage of the dictionary shall be made. In this regard, a conceptual design with information about the content and the methodology of the dictionary is now ready.

The Preparation phase A began before the grant of the project, if so. As for the Preparation phase B, it is included within the description and the realization of the dictionary, thus being a part of the three-year-period foreseen for the project.

Pilot study: although we started from a three year project also focused on the combinatorial potential of the noun, the main problem raised now is of a technical nature, that is, the evolution of the model from a bilingual to a multilingual dictionary.

[1] Project management: Prof. Dr. María José Domínguez Vázquez; Members of the project in the European Network of e-Lexicography: Mónica Mirazo Balsa, Prof. Dr. María Dolores Sánchez Palomino, Prof. Dr. Carlos Valcárcel Riveiro. Code: FFI2012-32456 (sponsored by Ministerio de Economía y Competitividad, Spain), 2013-2015. Key words: valency, multilingual, noun phrase, user and dictionary.

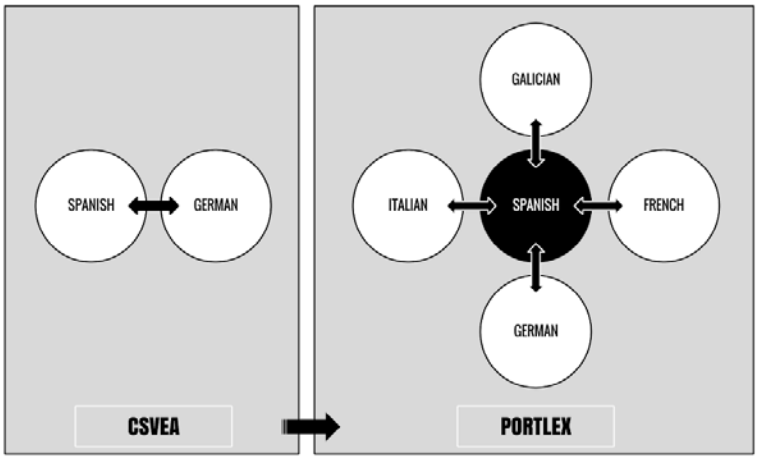

Image 1: *Evolution from CSVEA to PORTLEX*

Related to this are also the redevelopments of conceptional design, the microstructure of the entries, as well as the concept of meaning. The conceptional design is not static and all the work phases feedback.

2.2. Data acquisition

At this stage we have to acquire not only basic resources - like literature, articles etc... - but often computers facilities and other equipment.

By data acquisition we may also mean data extracted from online surveys relating to the type of users and their needs: it was therefore necessary to design surveys and pilot tests, as well as to extract and to analyse data. In our case, it may be said that it is the user who influences the model through the studies on dictionary use carried out with Spanish students studying German as a foreign language (Domínguez/Mirazo/Vidal, 2013) and with non-philologist Spanish students studying foreign languages (Domínguez/Valcárcel, 2014). The analysis of the obtained results led us to see the need to avoid labels, abbreviations and “too linguistic” terms and the advisability of making the presentation of entries more intuitive. This also explains the decision to include collaborative sections in our dictionary interface.

In a second phase is therefore required to continue the surveys and to rework, if applicable, the existing design of the architecture of networked information.

2.3. Computerisation

Our dictionary does not follow the procedure of automatic data extraction as we work with the following corpora:

- For Spanish: CREA (Corpus de Referencia del Español Actual[2]).
- For German: the German Reference Corpus (DeReKo), that is analysed via the query system COSMAS II.
- For French: FRANTEXT[3].
- For Galician: CORGA (Corpus de Referencia do Galego Actual[4]) and TILG (Tesouro Informatizado da Lingua Galega[5])
- For Italian: *Corpus dell'italiano scritto* (CORIS[6]), *itWaC*[7] and the database of the newspaper *La Repubblica*[8].

Even though several lists of recording abbreviations were developed for the starting database (CSVEA-Project), this new project foresees, to expand, to design and to implement multilingual corpus and a database management program. However, regarding the application to other languages, soon became apparent that the database that hosts the former CSVEA project could not be reused for PORTLEX. Many technical problems arose, so that colleagues and the software company advised us against enhancing the former database to other languages. The information search in CSVEA follows a modular structure in the query interface, so that the user can access different blocks of content depending on the extent of his search. This implies a substantial delay in planning, problems with the database structure and also a larger financial investment. Such searches can be made through the handling of different parameters that allow to display only the desired information. According to his search preferences, the user can thus obtain a complete view of the selected entry or, if desired, more specific or complex information. In other words, the display can only show the equivalent of a given noun or the complements that may accompany the same or a more complex search, like searching for all nouns sharing a complement with the same characteristics.

[2] <http://corpus.rae.es/creanet.html>

[3] <http://www.frantext.fr/>

[4] <http://corpus.cirp.es/corga/>

[5] <http://www.ti.usc.es/tilg/>

[6] <http://corpora.dslo.unibo.it/coris_ita.html>

[7] <http://wacky.sslmit.unibo.it>

[8] <http://sslmit.unibo.it/repubblica/>

2.4. Data processing and data analysis

When dealing with several languages, it is necessary to weigh the descriptive model and the possibility of including or excluding some information following theoretical-practical, applied and lexicographic criteria. This phase lasts a year, since it is carried out in parallel with the description of nouns. Experience shows that although based on a solid theoretical model, description and empirical work require certain restructurings. So this task and the analysis of nominal units necessarily overlap. Due to the enhancing of the model to other languages, it will be also necessary to take the inclusion of different contents into consideration during this phase.
Following are the steps taken:

1) **List of candidate entries to be selected**: Frecuency criteria based on the data obtained from CREA were applied for selecting the 200 most common Spanish nouns. After this first selection, a second filter was applied according to different criteria: the valential character of the nouns, the importance they may have from a contrastive point of view taking into account, among other factors, differences or similarities in the syntactic or the semantic valence or the existence of synonymy relations. The need to work with frequency criteria for Galician, German, Italian and French, despite being described as translation equivalents is due to the lack of a univocal relation between languages. Hence, among the possible equivalents, those showing the higher frequency rate are analysed first. In practice, the description is based on semantic fields (movement, cognition etc.).
The processes of elaboration and review are actually highly interconnected.

2) **Data processing and analysis**
a. Data processing: Descriptions of entries – Phase 1. This activity entails the following interrelated subtasks: i) To determine the number and definition of the meanings for each entry, according to the information contained in dictionaries and data resulting from corpora, ii) Selection of the examples from Corpora, iii) Description of the actantial schema for each meaning, as well of the most relevant combinatorial features at the syntactic and semantic levels, iv) In order to ensure, to the extent possible, the homogeneous treatment of the phenomena studied and the descriptive results, a manual for data input will be produced. This manual will also serve as methodological guide for new researchers who could join the project in the future. For theoretical and methodological reasons, it was intended that every working group included members of CSVEA project. The breakdown of tasks for each language is as follows [9].
b. Data analysis: Review and discussion of the entries of the Phase 1
c. Data processing: Development of the analysis of new dictionary entries**.** Phase 2.[10]
d. Data analysis: Review and discussion of the entries of the Phase 2
e. Data processing: Development of the contrastive modules: This working phase is divided into two subtasks: i) the contrastive module for the roman languages and ii) the contrastive module for the roman languages and German.
f. Data analysis: Review and discussion of the contrastive modules.
g. Finalization

2.5. Data analysis[11]

How do we work? i) To facilitate communication and collaboration among the members of the project, a working group on a social network has been opened. ii) Working meeting is also held every 15 days, since all project members work at Galician universities. iii) After the first phase is completed, project members will meet to solve possible difficulties encountered in analyzing noun phrases (Both phases of development and correction of new entries overlap in time).

2.6. Preparation for online release

Guided tour: Information and user guide on the database and the online dictionary

The aim here is to translate not only the theoretical but also methodological assumptions of the lexicographical work in a comprehensive and user-accessible format. It is also essential to give some simple guidelines on the operation of the database and the dictionary. This user guide is based on the data input manual. To this must be added the final review of the entries from a user perspective, as well of the content review, if necessary.

2.7. Afterlife

PORTLEX will be hosted on the server of the University of Santiago. However, more human resources are needed to ensure its updating, which brings us back to financing difficulties. Furthermore, users will have the option to send feedback to the dictionary team.

[9] **Analysis of German nouns** (Phase 1: 50%; Phase 2: 100%), **analysis of Spanish nouns** (Phase 1: 50%; Phase 2: 100%), **analysis of Italian nouns**: (Phase 1: 50%; Phase 2: 100%), **analysis of French nouns** : (Phase 1: 50%; Phase 2: 100%) and **analysis of Galician nouns** (Phase 1: 50%; Phase 2: 100%) Month 12: Phase 1: Analysis of nouns in different languages.

[10] A**nalysis of German nouns** (Phase 2: 100%), **analysis of Spanish nouns** (Phase 2: 100%); **analysis of Italian nouns**: (Phase 2: 100%), **analysis of French nouns**: (Phase 2: 100%) and **analysis of Galician nouns**: (Phase 2: 100). Month 25: Phase 2: Analysis of nouns in different languages.

[11] For further information on phases of the data analysis see 2.4.

2.8. Time span of the different phases

ACTIVITIES	PLANNING																																				
PREPARATION	*	FIRST YEAR												SECOND YEAR												THIRD YEAR											
A. Project conception, meetings, project application	▨																																				
B. Descriptive model		▨	▨	▨	▨	▨	▨	▨	▨	▨																											
DATA ACQUISITION	*	FIRST YEAR												SECOND YEAR												THIRD YEAR											
A. Sources and user surveys		▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																		
B. Data collection from corpora					▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																							
DATA PROCESSING 1	*	FIRST YEAR												SECOND YEAR												THIRD YEAR											
A. Selection of the unities to describe		▨	▨																																		
B. Phase 1: Dictionary entries																																					
a. Analysis of German nouns		▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																								
b. Analysis of Spanish nouns		▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																								
c. Analysis of Italian nouns				▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																								
d. Analysis of French nouns				▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																								
e. Analysis of Galician nouns				▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																								
C. Phase 2: Dictionary entries																																					
a. Analysis of German nouns														▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
b. Analysis of Spanish nouns														▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
c. Analysis of Italian nouns														▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
d. Analysis of French nouns														▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
e. Analysis of Galician nouns														▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
E. Elaboration of the contrastive modules																																					
Contrastive module among romance languages																								▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				
Contrastive module between romance languages and German																								▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				
DATA ANALYSIS	*	FIRST YEAR												SECOND YEAR												THIRD YEAR											
C. Revision and discussion of the descriptions of Phase 1.				▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																							
D. Revision and discussion of the descriptions of Phase 2.																		▨	▨	▨	▨	▨	▨	▨	▨	▨	▨										
E. Revision and discussion of contrastive modules. Finalization																																					
Contrastive Module among romance languages																										▨	▨	▨	▨	▨	▨						
Contrastive Module between roman languages and German																										▨	▨	▨	▨	▨	▨						
4. COMPUTERIZATION	*	FIRST YEAR												SECOND YEAR												THIRD YEAR											
Expansion, improvement, design and implementation of a multilingual corpus and management programs. Application to other languages		▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																								
6. PREPARATION FOR ONLINE RELEASE	*	FIRST YEAR												SECOND YEAR												THIRD YEAR											
Information and Management Guide for the database and the online dictionary																																	▨	▨	▨	▨	▨
AFTERLIFE																																					

* Before the grant project.

Table 1: *Process phases of the dictionary project and their time span*

This initial planning has undergone changes in the line with what was stated in paragraphs 2.1 and 2.3[12].

3. Summary of concerns and proposals

Within the framework of PORTLEX, a distinction should be drawn between “Considerations to take little account in the planning of electronic dictionaries”[13] and “Methodological, technical and human issues that affect the initial planning” [14]: The following chart gives an overview of different phenomena that can affect the initial planning:

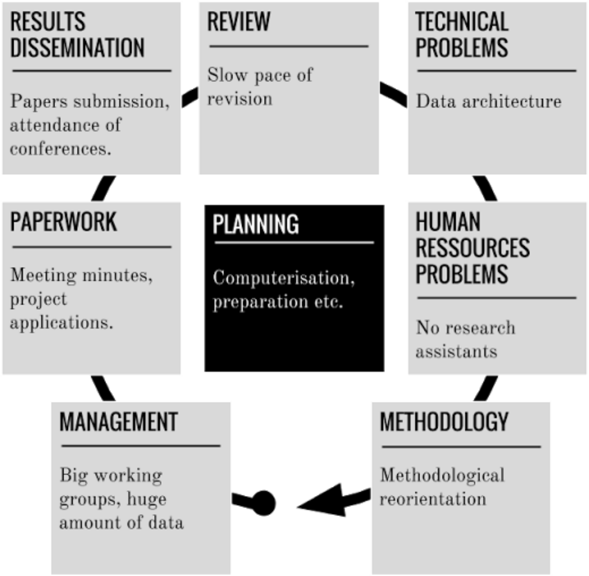
RESULTS DISSEMINATION
Papers submission, attendance of conferences.
REVIEW
Slow pace of revision
TECHNICAL PROBLEMS
Data architecture
PAPERWORK
Meeting minutes, project applications.
PLANNING
Computerisation, preparation etc.
HUMAN RESSOURCES PROBLEMS
No research assistants
MANAGEMENT
Big working groups, huge amount of data
METHODOLOGY
Methodological reorientation

Image 2: *Methodological, technical and human issues that affect the initial planning*

Within the initial planning, PORTLEX is not developed through a linear process, but an iterative process with feedback loops between the foreseen stages.

In other words, every new phase does not exactly start at the end of the previous one, since all phases do not involve linear processes, as initially stated in the planning. Phases mutually influence and modify each other during the development of every phase. In addition to this, the above factors affect the planning and undermine the initial planning. Therefore, these working processes must be taken into consideration for a more accurate and realistic planning.

[12] The following assumptions were made: i) **Reworking of the descriptive model on the premise of taking Spanish as the core language**. The possibility of such a multilingual dictionary was already proposed by Gouws (2014). This model of an axis dictionary with a core language (Spanish, in this particular case) as well as data collection from surveys, agrees with the idea of Tarp (2013: 304). A clear distinction should be drawn between the data recorded in the database and its visualization on display, ii) **Inclusion of new modules to ensure the attention to the needs of users. In this regard, the** initial descriptive model has also been changed. Although initially only meanings opening valence boxes were included for each noun, after weighing up different model nouns for the languages concerned, it was concluded that it was also relevant to include those meanings that, despite not having a valential complement, are commonly used in the targeted languages and thus relevant from a contrastive point of view. In this way, the dictionary may become a reference work for a wider range of audience .All this inevitably leads to a modification of the database structure. iii) In order to facilitate **user support**, it is also necessary to include a collaborative section in the dictionary interface. According to the diverse range of users, it is advisable to improve descriptive labels; iv) **The issue of re-using data from the former project (CSVEA)**. The inclusion of new languages entails reprocessing the data of CSVEA. However, this issue was not exhaustively covered in the initial planning and v) **Computerization issues** (see 2.3. for more details).

[13] For example a) Afterlife, b) Design, web Design (see: http://multimedia.ids-mannheim.de/mediawiki/web/images/7/7b/Programm_6_Arbeitstreffen.pdf), c) Empirical data on the dictionary (uses, users etc.).

[14] i) Concerns related to the **database** : a) New database structure, b) Re-using previous linguistic data; ii) Reworking of the **descriptive model**: a) Multilingual dictionary, b) Inclusion of new modules according to the needs of the user; iii) Concerns related to **human resources** (universities versus research centers)**:** a) In the case of Galician universities, where this project is being developed, it has become difficult to recruit external staff responsible for the database and technical issues. The impossibility of hiring qualified staff for computing tasks enforces project members, who certainly lack of training on this field, to solve these issues by themselves b) In addition, research tasks are also carried out by teachers. This implies that an increase in the teaching load, the assumption of new academic functions etc. could easily affect the dedication and the commitment of those involved in the current project c) The three-year period during which the project is being implemented prevents long-term training of young researchers; iv) **Financial concerns**: the lack of computer linguists led us to pay for the services of private consultants v) Management and coordination: another aspect that was not taken into account in the initial planning is project coordination and management (15 members), which entail more difficulties than initially expected; vi) **Errors in the initial planning**: Measures for the dissemination and presentation of the project in different forums and media were not included in the planning.

4. References

Domínguez Vázquez, M.J./Valcárcel Riveiro, C. (2014): Hábitos de uso de los diccionarios entre los estudiantes universitarios europeos: ¿nuevas tendencias?. In: Lexicografía de las lenguas románicas II. Aproximaciones a la lexicografía contemporánea y contrastiva. [Lexicography of the Romance Languages. Contemporary and Contrastive Lexicography] Ed. by Domínguez Vázquez, María José / Gómez Guinovart, Xavier / Valcárcel Riveiro, Carlos **(**forthcoming**),** Berlin: de Gruyter, (forthcoming).

Domínguez Vázquez, M. J./Mirazo Balsa, M./Valcárcel Riveiro, C. (2014): Evolución del diccionario bilingüe al multilingüe: de CSVEA a PORTLEX" In: Meliss, Meike / Sánchez Palomino, Mª Dolores / Sanmarco Bande, Mª T. (eds.): A lexicografía das linguas románicas: Estado da cuestión. München: Iudicium, (forthcoming).

Domínguez Vázquez, M. J. /Mirazo Balsa, M./Vidal Pérez, V. (2013): Wörterbuchbenutzung: Erwartungen und Bedürfnisse. Ergebnisse einer Umfrage bei Deutsch lernenden Hispanophonen,. In: Domínguez, M. J. (ed): Trends in der deutsch-sprachigen Lexikographie. Peter Lang: Frankfurt, 135-172.

Gouws, R. (2014): Towards bilingual dictionaries with Afrikaans and German as language pair. In**:** Domínguez Vázquez, M. J. / Mollica, F./ Nied, M. (2014): Zweisprachige Lexikographie zwischen Translation und Didaktik. (Lexikographica Series Maior), Berlin: de Gruyter, 15-28 **(**forthcoming**).**

Klosa, A.: Der lexikografische Prozess von elexiko (http://multimedia.ids-mannheim.de/mediawiki/web/images/7/76/Prozess_elexiko_Klosa.pdf).

Klosa, A. (forthcoming): 'The lexicographical process II: Online dictionaries'. In: *Dictionaries. An international encyclopedia of lexicography*. Supplementary volume: *Recent developments with special focus on computational lexicography*. Edited by Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard, Herbert Ernst Wiegand. Berlin/New York: de Gruyter.

Meyer, C.: Der lexikographische Prozess im deutschen Wiktionary (http://multimedia.ids-mannheim.de/mediawiki/web/index.php/4._Arbeitstreffen)

Tarp, S. (2013): El diccionario del futuro. In: Ruíz Miyares, L./ Álvarez Silva, M:R./ Muñoz Alvarado, A. (eds.): *Actualizaciones en Comunicación Social*. Santiago de Cuba: Centro de Lingüística Aplicada, vol. 1, 304-308.

Tiberius, C./ Schoonheim, T. (forthcoming): The Algemeen Nederlands Woordenboek (ANW) and its Lexicographical Process. In**:** Vera Hildenbrandt (ed.): Der lexikografische Prozess bei Internetwörterbüchern. 4. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie". Mannheim: Institut für Deutsche Sprache. (OPAL – Online publizierte Arbeiten zur Linguistik X/2014).