Advantages and Disadvantages in the Use of Internet as a Corpus: The Case of the Online Dictionaries of Spanish Valladolid-UVa[*]

Sven Tarp, *International Centre for Lexicography, Universidad de Valladolid, Spain; Department of Afrikaans and Dutch, University of Stellenbosch, South Africa* and *Centre for Lexicography, University of Aarhus, Aarhus, Denmark (st@asb.dk)*
and
Pedro A. Fuertes-Olivera, *Department of Afrikaans and Dutch, University of Stellenbosch, South Africa* and *International Centre for Lexicography, Universidad de Valladolid, Valladolid, Spain (pedro@emp.uva.es)*

Abstract: This paper initially discusses some of the consequences which the technological development has for lexicography, especially in terms of the different types of empirical basis which can be used in dictionary projects. The most important advantages and disadvantages of using the Internet as a corpus are then listed and compared to the usefulness of "traditional" corpora. As an example, the paper shows how the Internet is used as the main empirical source in order to select lemmata and meaning items in the Online Dictionaries of Spanish Valladolid-UVa. The methods and tools employed in the project are discussed together with the requirements to the lexicographers' competences, knowledge and skills. Finally, the paper provides some general conclusions as well as recommendations and hypotheses for future lexicographical work and research.

Keywords: INTERNET LEXICOGRAPHY, ONLINE LEXICOGRAPHY, LEXICOGRAPHICAL METHODOLOGY, EMPIRICAL BASIS, LEMMA SELECTION, MEANING SELECTION, LEXICOGRAPHICAL DATABASES, SPANISH DICTIONARIES, MONOLINGUAL DICTIONARIES, GENERAL DICTIONARIES

Opsomming: Voordele en nadele van die gebruik van die internet as 'n korpus: Die geval van die Aanlynwoordeboeke van Spaans Valladolid-UVa. Hierdie artikel bespreek aanvanklik sommige van die gevolge wat tegnologiese ontwikkeling vir die leksikografie inhou, veral in terme van die verskillende soorte empiriese basisse wat vir woordeboekprojekte gebruik kan word. Die belangrikste voordele en nadele van die gebruik van die internet as 'n korpus word dan gelys en vergelyk met die nuttigheid van "tradisionele" korpora. As voorbeeld toon die artikel hoe die internet as die belangrikste empiriese bron gebruik word om lemmata en betekenisitems vir die Aanlyn Woordeboeke van Spaans Valladolid-UVa uit te soek.

[*] This article was presented as a paper at the Twenty-first Annual International Conference of the African Association for Lexicography (AFRILEX), which was hosted by the Xitsonga and Sesotho sa Leboa National Lexicography Units, Tzaneen, South Africa, 4–6 July 2016.

Die metodes en werktuie wat in die projek gebruik word, word bespreek, sowel as die vereistes wat aan die leksikograwe se bevoegdhede, kennis en vaardighede gestel word. Ten slotte verskaf die artikel 'n paar algemene gevolgtrekkings, asook aanbevelings en hipoteses rakende leksikografiese werk en navorsing.

Sleutelwoorde: INTERNETLEKSIKOGRAFIE, AANLYN LEKSIKOGRAFIE, LEKSIKOGRAFIESE METODOLOGIE, EMPIRIESE BASIS, LEMMASELEKSIE, BETEKENISSELEKSIE, LEKSIKOGRAFIESE DATABASISSE, SPAANSE WOORDEBOEKE, EENTALIGE WOORDEBOEKE, ALGEMENE WOORDEBOEKE

1. Introduction

Looking at the overall dictionary compilation process as described by Fuertes-Olivera and Tarp (2014: 85), there are three instances where lexicographers may need access to empirical data in order to do a good job. The first instance is when they are looking for information about the foreseen users' lexicographical needs with a view of preparing a dictionary concept which can assist these users in solving their needs. The second one is when they are selecting and preparing the lexicographical data to be included in the dictionary. And the third one is when they evaluate the usefulness of the dictionary in terms of user satisfaction. In fact, there is also a fourth situation where external empirical data may be required, i.e. when analysing the market in order to determine the sales possibilities of the product, but this is more related to the business side of the project than to lexicographical aspects in the narrow sense of the word. Anyway, in each of these situations there is a set of methods that may appear to be more or less appropriate, i.e. more or less reliable and fast in terms of both productivity and quality of the final product.

In the following, we will look at the empirical bases and the corresponding methods that can be applied when selecting lemmata and meaning items (senses) in a lexicographical online project. We will then discuss some of the most important advantages and disadvantages when using the Internet directly as a corpus, and compare them to the usefulness of "traditional" text corpora. As an example we will take a project currently carried out at the International Centre of Lexicography at the University of Valladolid, namely the *Diccionarios en Línea de Español "Universidad de Valladolid"*, in the following referred to as the *Online Dictionaries of Spanish Valladolid-UVa*. The project, which was originally initiated as a collaboration between the Valladolid-based Centre and its sister Centre for Lexicography at Aarhus University, is based on the lexicographical function theory and inspired by a similar Danish project; cf. Fuertes-Olivera and Bergenholtz (2015). Finally, we will present the hitherto experience and introduce the need to count on intuition as an intangible but highly relevant and unavoidable method in dictionary making.

2. Relationship between lexicography and technology

In a historical perspective, a both intimate and complex relationship can be observed between lexicography and technology. This implies, among other things, that technological development may lead not only to new tools with which lexicographers can perform their art and craft, but also to new empirical bases from which they can retrieve their data as well as the need for, and possibility of, developing new methods that can be applied in this respect. The reflection is especially relevant in historical periods as the present one where new disruptive technologies are being introduced in lexicography with consequences that can still not be completely grasped:

> Today we are in the middle of a new transition of the material and technological basis of lexicography with the introduction of new production tools and methods as well as new platforms and media for presenting the lexicographic product and the extensive use of corpora for the collection of material. The development and technological innovation are going faster than ever before. (...) We know the point of departure but we still only have a vague idea of where we will eventually arrive. (Gouws and Tarp 2016)

Generally, there is a variety of sources from which lexicographers can obtain their data. Bergenholtz and Tarp (1995: 90-96) discuss, among the most important, introspection, multispection, external experts, existing dictionaries, handbooks, textbooks, example cards, and text corpora. With the exception of the three former, these empirical sources have only been possible thanks to the technological development at its various stages: the invention of paper, pens, bookbinding, printing machines, computers, and databases. Since then, with the introduction and development of the web technology, another empirical data source has been put at the disposal of lexicography, namely the Internet.

It is interesting to note that Bergenholtz and Tarp (1995) discuss the consultation of external experts as a form of multispection and, implicitly, the use of one's own knowledge as a form of introspection. It may be so, but it seems nonetheless that there is a difference between the use of introspection in terms of language skills and competence, as it is normally understood within linguistics, and the use of up-to-date expert knowledge stored in someone's memory. As Tarp (2008: 131-136) has argued, in the preparation of various types of dictionary it is important to distinguish between language skills and learned knowledge of a given language, for instance as it is provided by linguistic theory. In this respect, it seems reasonable also to distinguish between the use of language competence and the use of expert knowledge in the dictionary compilation process. Hence, although there is certain terminological confusion in the existing lexicographical literature, introspection — rather than an empirical basis in itself — should be considered a method to obtain specific types of empirical data. It is a method to "look" into oneself in order to retrieve material for different purposes. The "internal" empirical bases on which lexicographers can draw by means of this method are language competences, skills, and knowledge, to which can be added personal experience in general.

The various empirical sources are seldom used alone. In a book review, Kilgarriff (2012) provides an example on how two different types of empirical basis are combined:

> I noticed a lexicographical bloomer. On pp 211-213 we have an analysis of the English phrasal verb *call back*. It is given six meanings of which the sixth is given the example "I cannot call his face back." As an English native speaker, I go *eeeeeugh*. This is blazingly wrong. (We might say "I cannot recall his face.") A little research revealed that this 'example sentence' exists in a number of dictionaries and translation tools: a dictionary error that has been copied and recopied from dictionary to dictionary. (Kilgarriff 2012: 28)

When Kilgarriff says he "goes eeeeeugh" as a native speaker of English, this suggests that his mother-tongue competence warns him about a possible problem which he subsequently confirms and explains through the consultation of other empirical bases, in this case existing dictionaries and translation tools. This is obviously the right method to apply in such cases because it implies that "the lexicographer's primary source of evidence for how a word behaves switches from subjective to objective; from introspection to looking at contexts" (Kilgarriff 1997: 111). It should be noted that Kilgarriff here speaks about the *primary* source of evidence, not the only one, although he errs when defining introspection and "looking at contexts" as sources, inasmuch as they are both methods to access the real sources of evidence.

3. Corpus versus Internet: Preliminary discussion

The first electronic text corpora were introduced in the 1960's, and since then they have never stopped growing. The two first decades after their appearance were characterized by a fierce battle of ideas between the researchers who defended the relevance of corpora for both linguistics and lexicography, and those who opposed this idea with various arguments, generally in favour of introspection as a much more appropriate method to get empirical material. One of the defenders of introspection was Lees (1962) who declared straightforwardly:

> You are a native speaker of English; in ten minutes you can produce more illustrations of any point in English grammar than you will find in many millions of words of random text. (Lees 1962: 110)

Little by little the discussion faded out. Half a century after their introduction, there is no longer any doubt that electronic text corpora can be of great value not only to linguistic research but also to lexicographers when performing a series of tasks in connection with the compilation of dictionaries. This has been argued by various scholars engaged in practical lexicography, among them Bergenholtz (1996), Atkins and Rundell (2008), and Hanks (2012). The proof of the pudding is the existence of many high-quality dictionaries which have been compiled based upon this type of empirical basis (see e.g. Sinclair 1997), although the eagerness has sometimes gone too far, none at least in connection with the selection of terms and definitions in specialized dictionaries; cf. Tarp (2016), and Xue and Tarp (2016).

However, a negative consequence of this generally positive development is that introspection as a method to make use of one's own competences and knowledge is occasionally underestimated or even ignored. Although "lexicographers should never rely solely on the introspective approach" (Bergenholtz and Tarp 1995: 92), especially in cases of doubt, it is frequently forgotten that introspection always lays as a filter at the bottom of the lexicographer's choices inasmuch as no seriously working dictionary maker would introduce linguistic or any other type of data with which he or she disagrees — says "eeeeeugh" in Kilgarriff's expressive but very accurate terminology — without first negotiating their correctness with other empirical sources.

Today, corpora composed of texts containing hundreds of millions of words are available to the compilers of dictionaries. In this respect, Big Data is already a reality, but the understandable excitement created by this development should never be allowed to overshadow the fact that no corpus, however big, can stand up to the enormous collection of texts and words which can be accessed through the Internet. The development of methods allowing for the use of this almost unlimited empirical basis constitutes undoubtedly a challenge more and more relevant to lexicography.

According to Fuertes-Olivera (2012: 51), a *lexicographical corpus*, i.e. a corpus that can be used to assist dictionary making, can be defined as "any collection of texts where lexicographers can find inspiration for completing the dictionary structures they need when making a real dictionary". As already mentioned, the Internet is made up by a collection of texts. Thus, if a lexicographer can find inspiration in this big collection of texts, the Internet can also be considered a type of lexicographical corpus according to the above definition. This is also the point of view of Kilgarriff and Grefenstette (2003: 334) who write that "the answer to the question 'Is the web a corpus?' is yes."

In this respect, there are two different ways of using the Internet in relation to a lexicographical project, namely 1) constructing a corpus of texts found on the Internet, and 2) using the Internet directly as a corpus, in both cases by means of search engines and other tools. Each of these two types of lexicographical corpus has its advantages and disadvantages. Below are listed some of the advantages when the Internet is used directly as a corpus, in comparison to the use of "traditional" corpora that are made up of collections of texts, whether or not these texts are taken from the Internet or elsewhere:

— The lexicographers have access to many more texts than the ones included in any corpus of selected texts.

— The texts are always up-to-date.

— Time and money are saved when it is not necessary to compose a separate corpus (which is a requirement in relation to specific types of dictionaries, in particular specialized ones).

— The search process can easily be limited to specific geographic areas, a fact that is especially important for a multinational language as Spanish.

— The use of the Internet may lead to the identification and selection of more meaning units than those that can be found in a separate corpus.

As to the disadvantages when using the Internet directly as a corpus, the following seem to be the most important:

— The quality and origin of the texts cannot be controlled.

— The authors of some of the texts may not be real persons.

— The authors may have a low proficiency level in the language in question.

— The texts may not have been revised and corrected.

— It is difficult to calculate the frequency of the linguistic phenomena appearing in the texts.

Some of the above disadvantages may not be relevant to concrete dictionary projects. Gudmann (2014: 32), for instance, argues that "information about frequency (…) is not particularly relevant to a general monolingual reception dictionary". In other cases, the disadvantages can be neutralized, or at least considerably reduced, by a well-trained lexicographer who plays an *active role* based on his or her language competence, skills, knowledge and experience. We will return to this question later on. At this point, our preliminary conclusion is that in spite of the undeniable disadvantages, it is perfectly possible, and even beneficial, to use the Internet as the main empirical source, without resorting to the "traditional" text corpora, when the objective is the production of dictionaries of still higher quality.

4. Selecting lemmata

We sincerely doubt that the traditional corpus composed of a collection of texts is the most appropriate empirical source for lemma selection, especially if this selection has to be done from scratch. To the best of our knowledge, the big general dictionaries that use corpora for this purpose are mostly dictionaries that had their basic lemma stock selected before the introduction of corpora which are now "only" used to provide additional lemmata, among other data. A different method and empirical basis are therefore required when the challenge is a quick and reliable selection of lemmata to a completely new lexicographical project of the magnitude of the Online Dictionaries of Spanish Valladolid-UVa which are planned to handle more than one hundred thousand lemmata and many more senses. The primary empirical basis chosen for this project was therefore the Internet accompanied by a method which will be described in this section.

The basic idea is that the Internet already contains a considerable number of smaller or bigger word lists for free access and use. The challenge is therefore to find these lists and make use of them. This is done by means of an Internet crawler that has been specially designed for this purpose by the Danish company Ordbogen.com which, due to its business model, is the world's currently most successful provider of online dictionaries in subscription.

Once a number of useful word lists have been found by the Internet crawler, these lists are copied and pasted into a so-called *lemma loader* (see Figure 1), another tool developed by Ordbogen.com and conceived by Professor Emeritus Henning Bergenholtz from the Centre for Lexicography in Aarhus. The lemma loader assigns automatically a lemma to a card in the database and has the advantage that it does not reduplicate the lemmata, but rejects them if they are already stored in the database.

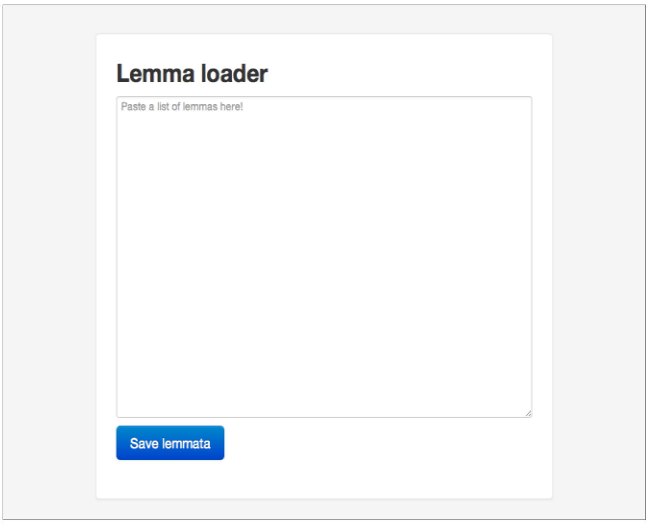


Figure 1: Screenshot of the lemma loader with the field where the copied wordlists are pasted

The experience shows that this method to select lemmata by means of an Internet crawler and a lemma loader is a very efficient, fast and totally reliable method in the case of one-word lemmata. This is reflected by the fact that only one month after the lemma selection to the Online Dictionaries of Spanish Valladolid-UVa started (July 2013), the database contained already 58.000 cards with one-word lemmata.

The next step in the lemma selection process is a manual revision which takes place when the formal grammatical data are attached to the lemma card. The revision is performed by the editor-in-chief (Pedro A. Fuertes-Olivera). Due to the characteristics of the dictionaries and the almost unlimited storage capacity of the database, the project does not work with lemma inclusion criteria but only with exclusion criteria. With this approach, only lemmata that clearly represent spelling mistakes or cannot be documented in the empirical basis, i.e. the Internet, are excluded, even when the latter can be found in some old word lists but not on the Internet as such. Until now there have only been few cases where a lemma had to be excluded, a fact which also points to the efficiency of the method.

In order to guarantee a systematic treatment of the language, after the initial selection a number of thematic lists containing colours, numbers, cities of a certain size, rivers of more than 1.000 km, etc. were elaborated and the corresponding words introduced in the database as lemmata. This work lasted from September to November 2013 and resulted in the inclusion of additional 10.000 lemmata.

Apart from the mentioned empirical sources, other sources are also used in order to provide a flow of new lemmata. For instance, when the lexicographers are working on the Internet in order to identify meaning items to the selected lemmata (see Section 5), they simultaneously detect a considerable number of synonyms, antonyms and word combinations which are continuously introduced into the database as new lemmata.

Finally, there is the question of idioms and other fixed expressions which are generally selected as lemmata in their own right in the Online Dictionaries of Spanish Valladolid-UVa. Here there are four sources: 1) When the lexicographers are detecting meaning items they occasionally come up with such fixed expressions which are sent the editor-in-chief who then evaluates and analyses them by googling on the Internet. 2) Existing dictionaries are also used as sources in this respect, and 3) the same is the CREA Corpus composed and published by the Royal Spanish Academy for free use. 4) Finally, a number of fixed expressions are also found in other sources, e.g. books and articles read by the lexicographers in connection with other tasks.

As can be seen, it is only the three last sources of fixed expressions where the Internet is not the empirical basis for the selection of lemmata for the Online Dictionaries of Spanish Valladolid-UVa. Generally, the overall process can be characterized as a very fast, efficient and low-cost process which, until now, has resulted in about 20% more lemmata than those contained in the hitherto biggest Spanish dictionaries.

5. Selecting meaning items

The method developed to select meaning items in the Online Dictionaries of Spanish Valladolid-UVa is strongly inspired by a similar method used in the Danish Internet Dictionaries (see Bergenholtz and Agerbo 2014), but it also has some particularities of its own. Roughly speaking, the meaning selection method encompasses the following 15 steps or actions:

1. A lemma contained in the database is chosen in the lexicographer's user interface (see Figure 2)

2. The button "Google" to the left in the lexicographer's interface is activated.

3. A "traditional" Google-search result appears (See Figure 3).

4. The first (3-20) pages are skipped because they only contain lexicographically irrelevant data.

5. The minitexts appearing on each page are read in order to get a general idea of what it is all about.

6. Using the method "copy and paste" the relevant parts of the minitexts are copied into a Word document.

7. Simultaneously, collocations, examples, synonyms, antonyms and word formations are selected in order to be introduced into the respective fields in card representing the sense in question in the lexicographer's interface (see Figure 4 and 5). Idioms and fixed expressions are sent to the editor-in-chief for further evaluation.

8. A number of Google pages are reviewed until no more new data appear and everything is repeated. The number of pages depends on the characteristics of each lemma as well as the lexicographer's intuition based upon experience.

9. Once a satisfactory amount of empirical data has been selected, these data are grouped according to meaning.

10. Based on the groups of data the first definitions are written according to the lexicographical instructions prepared by the editor-in-chief.

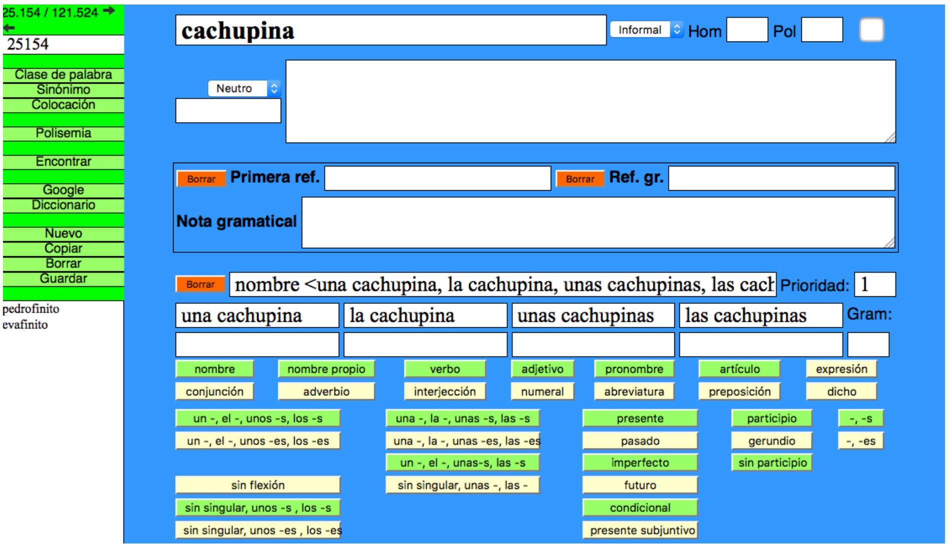
11. Now the lexicographer decides whether he or she is satisfied, or if it is necessary to repeat the process, or part of the process, in order to obtain a satisfactory amount of empirical evidence.

12. When the lexicographer has finished meaning selection and written the definitions of the senses addressed to a lemma, a message is sent to the editor-in-chief.

13. The editor-in-chief revises the definitions and compares them with the ones appearing in four Spanish dictionaries (see Section 6). If something is missing, this may lead to a new search process as it is a basic principle in the project that no definition is copied from other dictionaries.

14. If the definitions are related to specialised terms appearing in general language, external experts may be consulted in order to control their correctness.

15. When the editor-in-chief is satisfied — and other relevant data such as grammar, synonyms, antonyms, word formations, collocations and example sentences have been included — the lemma in question is indicated for online publication.



25.154 / 121.524

25154

cachupina

Informal

Neutro

nombre <una cachupina, la cachupina, unas cachupinas, las cach

Prioridad: 1

una cachupina

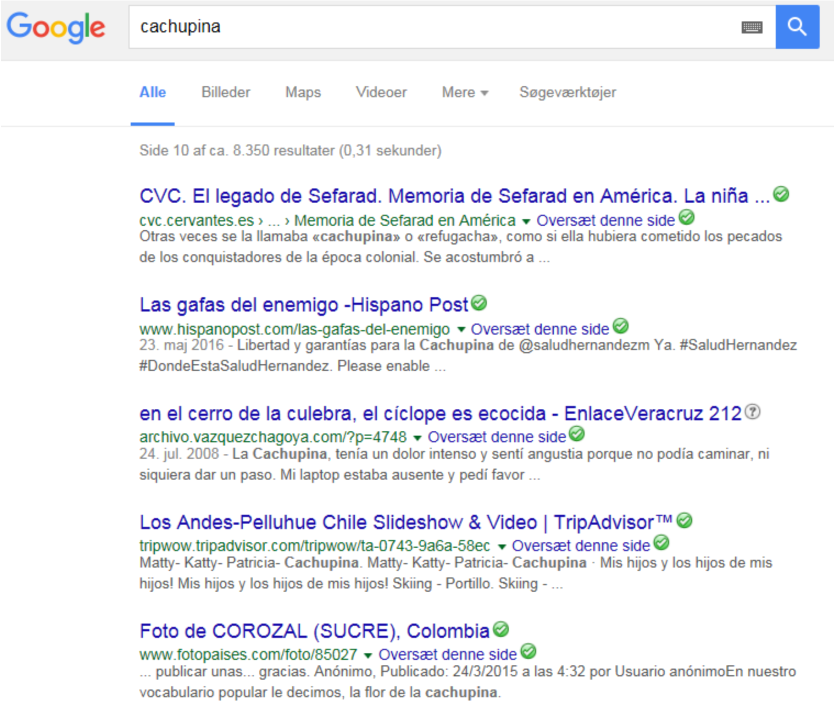
la cachupina

unas cachupinas

las cachupinas

pedrofinito
evafinito

Figure 2: The lexicographer's interface for grammar to the word "cachupina"



cachupina

Side 10 af ca. 8.350 resultater (0,31 sekunder)

CVC. El legado de Sefarad. Memoria de Sefarad en América. La niña ...
cvc.cervantes.es › ... › Memoria de Sefarad en América ▾ Oversæt denne side
Otras veces se la llamaba «**cachupina**» o «refugacha», como si ella hubiera cometido los pecados de los conquistadores de la época colonial. Se acostumbró a ...

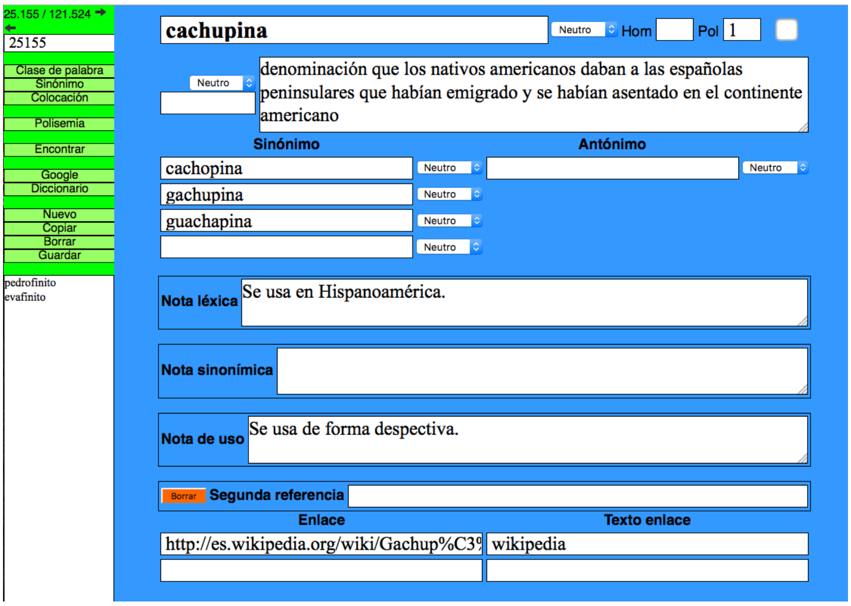
Las gafas del enemigo -Hispano Post
www.hispanopost.com/las-gafas-del-enemigo ▾ Oversæt denne side
23. maj 2016 - Libertad y garantías para la **Cachupina** de @saludhernandezm Ya. #SaludHernandez #DondeEstaSaludHernandez. Please enable ...

en el cerro de la culebra, el cíclope es ecocida - EnlaceVeracruz 212
archivo.vazquezchagoya.com/?p=4748 ▾ Oversæt denne side
24. jul. 2008 - La **Cachupina**, tenía un dolor intenso y sentí angustia porque no podía caminar, ni siquiera dar un paso. Mi laptop estaba ausente y pedí favor ...

Los Andes-Pelluhue Chile Slideshow & Video | TripAdvisor™
tripwow.tripadvisor.com/tripwow/ta-0743-9a6a-58ec ▾ Oversæt denne side
Matty- Katty- Patricia- **Cachupina**. Matty- Katty- Patricia- **Cachupina** · Mis hijos y los hijos de mis hijos! Mis hijos y los hijos de mis hijos! Skiing - Portillo. Skiing - ...

Foto de COROZAL (SUCRE), Colombia
www.fotopaises.com/foto/85027 ▾ Oversæt denne side
... publicar unas... gracias. Anónimo, Publicado: 24/3/2015 a las 4:32 por Usuario anónimoEn nuestro vocabulario popular le decimos, la flor de la **cachupina**.

Figure 3: Result of Google search for the word "cachupina"



25155

pedrofinito
evafinito

cachupina — Neutro — Hom — Pol 1

Neutro — denominación que los nativos americanos daban a las españolas peninsulares que habían emigrado y se habían asentado en el continente americano

Sinónimo		Antónimo	
cachopina	Neutro		Neutro
gachupina	Neutro		
guachapina	Neutro		
	Neutro		

Nota léxica: Se usa en Hispanoamérica.

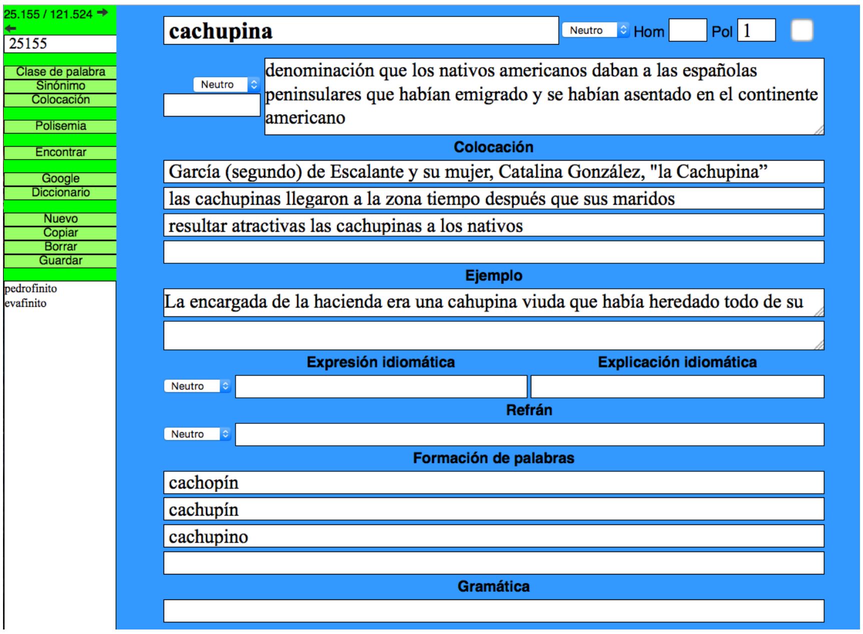
Nota sinonímica:

Nota de uso: Se usa de forma despectiva.

Segunda referencia:

Enlace	Texto enlace
http://es.wikipedia.org/wiki/Gachup%C3%	wikipedia

Figure 4: The lexicographer's interface for the introduction of definition, synonyms and antonyms to the word "cachupina"



25155

pedrofinito
evafinito

cachupina — Neutro — Hom — Pol 1

Neutro — denominación que los nativos americanos daban a las españolas peninsulares que habían emigrado y se habían asentado en el continente americano

Colocación

- García (segundo) de Escalante y su mujer, Catalina González, "la Cachupina"
- las cachupinas llegaron a la zona tiempo después que sus maridos
- resultar atractivas las cachupinas a los nativos

Ejemplo

La encargada de la hacienda era una cahupina viuda que había heredado todo de su

	Expresión idiomática	Explicación idiomática
Neutro		

Refrán: Neutro

Formación de palabras

- cachopín
- cachupín
- cachupino

Gramática

Figure 5: The lexicographer's interface for the introduction of collocations, example sentences, fixed expressions and word formations to the word "cachupina"

Currently (October 2016), the database of the Online Dictionaries of Spanish Valladolid-UVa contains about 55,000 finished cards (each of them representing one sense) that are ready for publication. The experience until now shows that for 70 percent of the lemmata it is sufficient to work with the minitexts that appear as a result of the Google search. For the remaining 30 percent it is therefore necessary to activate one or more of the links in order to find additional data in the unfolded documents. In this last case, the required data can be found 90 percent of the time. Only in 10 percent of the cases, representing 3 percent of the totality of lemmata, is it necessary to perform a new search with a variant of the lemma in question. This means that for 97 percent of all lemmata one Google search is sufficient to obtain the empirical material required to select meaning items and write definitions of the desired standard.

6. Comparison with similar Spanish dictionaries

As mentioned above, after writing the definitions of the different senses, these senses are compared with those found in four Spanish dictionaries, namely:

— María Moliner: *Diccionario de Uso del Español* (DUE)

— Aquilino Sánchez Pérez: Gran *Diccionario de Uso del Español Actual* (GDUEA)

— Manuel Seco: *Diccionario del Español Actual* (DEA)

— Real Academia Española: *Diccionario de la Lengua Española* (DLE)

These four dictionaries are among the biggest and most prestigious general dictionaries of Spanish. In this respect, the comparison also serves as a sort of quality control and indication of what could be improved (see action 13 in Section 5). The comparison has so far been favourable to the Online Dictionaries of Spanish Valladolid-UVa, as it shows that each lemma treated with the described method has an average of 30-40 percent more senses than the ones found in the four other dictionaries.

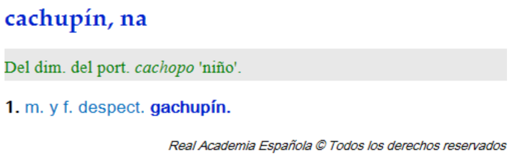
Table 1 shows the number of senses which the five mentioned dictionaries provide to eight different lemmata. It is in no way representative but just an indication of how the method described in the previous Section in some cases can generate a bigger number of senses.

Word	DUE	GDUEA	DEA	DLE	Valladolid
ababol	1	1	1	2	3
cabila	1	1	2	2	3
cable	4	4	4	6	11
cabestro	4	4	2	4	6
eclipsar	2	2	2	2	3

eclipsarse	2	2	2	3	4
halagar	4	2	2	4	3
machaca	5	5	5	6	15

Table 1: Comparison between five Spanish dictionaries in terms of number of senses addressed to eight selected lemmata

The tendency reflected in Table 1 is corroborated by Gudmann (2015) who has studied five Spanish online dictionaries, among them the one published by the Royal Spanish Academy, and identified a surprisingly big number of meaning lacunae. Another illustration of this phenomenon is the treatment of *cachupín* and *cachupina* which are two words used in some parts of Latin America and presented as one and the same lemma in the four other dictionaries mentioned above, each of them with only one sense. Figure 7 shows how they are handled by the Real Spanish Academy in the online version of its *Diccionario de la Lengua Española*:



cachupín, na

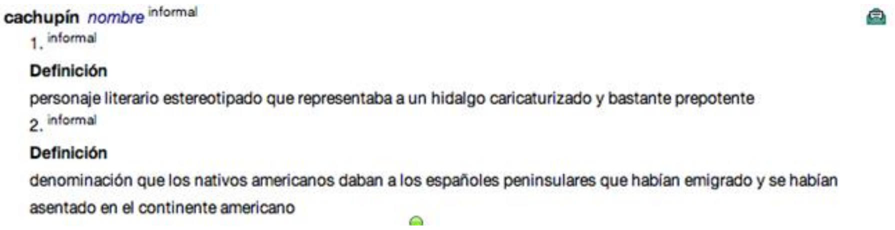
Del dim. del port. *cachopo* 'niño'.

1. m. y f. despect. **gachupín.**

Real Academia Española © Todos los derechos reservados

Figure 6: The lemma "cachupín, na" in the *Diccionario de la Lengua Española*

This way of presenting the two words can rightly be considered sexist, as if the feminine word *cachupina* was just subordinated to the masculine word *cachupín*. In the Online Dictionaries of Spanish Valladolid-UVa, the two words are therefore treated separately and listed as two different lemmata. This also suggests that a separate Google search for meaning items has been performed for each of the two words. The surprising result of this method is that the masculine word now appears with two senses while the feminine word includes no less than eight senses, i.e. a total of ten senses, as can be seen in the beta screenshots of the reception dictionary extracted from the database and presented in Figure 7 and 8.



cachupín *nombre* informal

1. informal

Definición

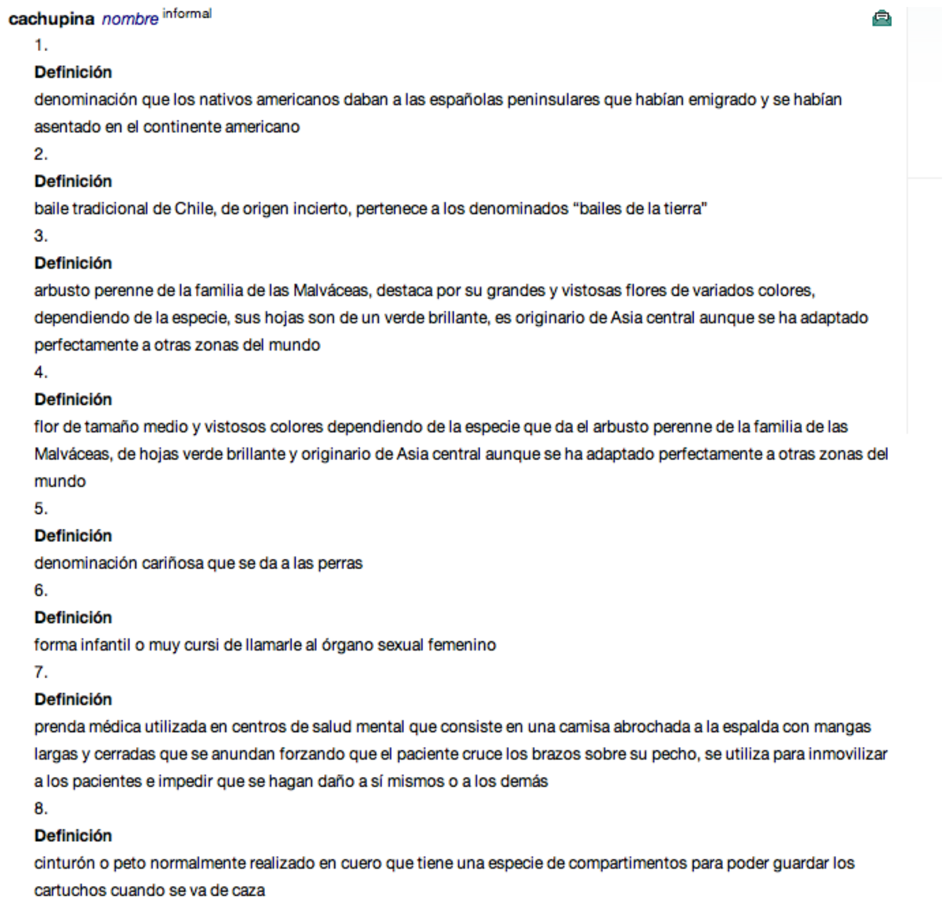
personaje literario estereotipado que representaba a un hidalgo caricaturizado y bastante prepotente

2. informal

Definición

denominación que los nativos americanos daban a los españoles peninsulares que habían emigrado y se habían asentado en el continente americano

Figure 7: The lemma "cachupín" in the Online Dictionaries of Spanish Valladolid-UVa



cachupina *nombre* informal

1.
Definición
denominación que los nativos americanos daban a las españolas peninsulares que habían emigrado y se habían asentado en el continente americano

2.
Definición
baile tradicional de Chile, de origen incierto, pertenece a los denominados "bailes de la tierra"

3.
Definición
arbusto perenne de la familia de las Malváceas, destaca por su grandes y vistosas flores de variados colores, dependiendo de la especie, sus hojas son de un verde brillante, es originario de Asia central aunque se ha adaptado perfectamente a otras zonas del mundo

4.
Definición
flor de tamaño medio y vistosos colores dependiendo de la especie que da el arbusto perenne de la familia de las Malváceas, de hojas verde brillante y originario de Asia central aunque se ha adaptado perfectamente a otras zonas del mundo

5.
Definición
denominación cariñosa que se da a las perras

6.
Definición
forma infantil o muy cursi de llamarle al órgano sexual femenino

7.
Definición
prenda médica utilizada en centros de salud mental que consiste en una camisa abrochada a la espalda con mangas largas y cerradas que se anundan forzando que el paciente cruce los brazos sobre su pecho, se utiliza para inmovilizar a los pacientes e impedir que se hagan daño a sí mismos o a los demás

8.
Definición
cinturón o peto normalmente realizado en cuero que tiene una especie de compartimentos para poder guardar los cartuchos cuando se va de caza

Figure 8: The lemma "cachupina" in the Online Dictionaries of Spanish Valladolid-UVa

The above differences should not be absolutized as the number of lemmata and senses collected in the database of the Online Dictionaries of Spanish Valladolid-UVa cannot be compared directly with the four other dictionaries of Spanish. There are two reasons for this. The first one is that the latter four are all printed dictionaries which suffer from the well-known space restraints. And the second one is that the "paper philosophy" continues to influence current Spanish lexicography even when it goes online. In this respect, a serious problem is that the selection criteria have still not been adapted to the new technology. Rundell (2015) shows in a condensed way how these criteria have been turned upside down by recent developments:

> So when there are no space constraints, it may make sense to turn the question around and — rather than asking 'does this word pass my inclusion tests?' — we

> should ask instead 'are there good reasons for not including this word'? (Rundell 2015: 312)

This change in selection criteria suggests that lexicographers, who only a few decades ago should justify the inclusion of any new lemma or sense because it frequently meant the exclusion of other data, are now challenged with the need to justify the non-inclusion of new lemmata and senses. In this respect, we do not know how many unpublished senses the other Spanish dictionaries may have in their databases. And neither do we know how many additional meaning items could be identified in their corpora if it was required. The comparison between the five dictionaries presented in Section 6 should therefore be taken as an indication of the new possibilities which the use of the Internet put at the disposal of lexicography, and not as an evidence-based fact that accounts for all aspects. At the end of the day what should be compared are not the dictionaries as such but the methods and empirical bases used to provide their lexicographical data.

With this in mind, our **current hypothesis**, which should be subjected to further research, is that *a corpus composed of selected texts is more appropriate to identify the most typical and frequent words, senses and behaviours of words, whereas the use of the Internet directly as a corpus is a more appropriate method when it is a question of detecting the less typical and frequent words, expressions, senses and behaviours of words.*

7. Quality and productivity

From the previous discussion it follows that the use of the Internet directly as a corpus represents a promising method to exploit two relevant results of recent technological development, namely 1) that the Internet today comprises an almost "unlimited" number of texts and words, and 2) that a modern digital dictionary is sustained by a database with almost "unlimited" storage capacity. The choice of method is therefore both an interesting topic for academic discussions and a question with big practical and economic consequences. Today the challenge for publishing houses and lexicographical teams in general is not only to compile high-quality dictionaries but also to guarantee high productivity in the compilation process. An increasing number of lexicographers are becoming aware that their discipline is submerged in a crisis which in a certain manner could be described as a struggle between life and death. This crisis is determined by two opposed tendencies in current lexicography: On the one hand, many publishers of high-quality dictionaries are closing down their dictionary departments due to lack of income and a sustainable business model. On the other hand, an increasing number of free-access dictionaries of dubious quality are placed on the Internet by generally well-intended but insufficiently trained people.

The results of this development are many and mostly negative. Although dictionaries, due to the online media, have more users than ever before, many of these users experience problems when they try to make use of the information retrieved from such dubious dictionaries, a fact which many language and translation teachers will recognize. The unavoidable result is that an increasing number of users, who are conscious of their needs, turn their back to lexicography and look in other information tools for assistance. Some big Spanish enterprises that are willing to pay for the service are, for instance, critical of the current standard of Spanish online dictionaries, leaving a market open for projects like the Valladolid-UVa. However, the problem is further exacerbated by the fact that many modern users of the Internet, especially young people, expect the service to be free as discussed by Gouws and Tarp (2016).

Publishing houses all over the world are struggling to find a solution to these challenges which are very complex. One of the necessary counter-measures to the present crisis is undoubtedly to *increase productivity* in the dictionary compilation process with a view to reducing costs and finding a sustainable business model, i.e. increase productivity without compromising quality. Productivity has many faces and can only be increased through the *integration of user-friendly technology, efficient methods and well-trained and motivated lexicographers*. In modern dictionary projects as the Valladolid-UVa, the lexicographer's interface is the central working tool by means of which the lexicographers introduce their data into the database. In this respect, Tarp (2015) writes:

> The lexicographer's interface is basically a *means of production*. It should therefore be designed with a view to guaranteeing both high productivity and the highest possible quality of the resulting product, i.e. the data stored in the database. This requires above all that it contains all the fields needed to introduce lexicographical data of the foreseen types into the database. But it is also important that the interface is as user-friendly as possible in order to facilitate the lexicographer's job, reduce the number of mistakes, economise on the resources employed, and shorten the total production time. (Tarp 2015: 234)

Figures 2, 4 and 5 in Section 6 reproduce three screenshots of the lexicographer's interface related to the lemma *cachupina*. The first one represents the mother card in which the grammatical data common to all senses of the lemma are introduced. The second and third screenshots show two pages of the card representing the first sense of the lemma. The pages are structured according to the different tasks that have to be performed and in such a way that the need to swift between one and another by means of the functional buttons to the left is reduced to a minimum. The main idea is that the lexicographers should feel comfortable when they work with this interface. Its user-friendly design together with the methods described in the previous sections is the precondition for the high productivity that characterizes the compilation of the Online Dictionaries of Spanish Valladolid-UVa.

The identification of meaning items and the introduction of definitions and other lexicographical data into the database started in March 2014. The experience shows that a lexicographer can finish an average of 4 to 6 senses per hour with the described method and production tool. The experience also shows that productivity decreases after four or five hours as the job demands a high degree of concentration, for which reason the four lexicographers doing this part of the job only work four hours daily on the project. But if we make an abstraction from this fact, it would mean that a full-time lexicographer in an 8-hour work day would be able to finish about 40 senses and in a 40-hour week about 200 senses, which adds up to a total of about 9.000 senses per lexicographer in a 45-week work year.

The overall outcome is that the four half-time lexicographers attached to the project have finished a total of about 40.000 sensed (cards) from March 2014 to July 2016. This could be compared with the small army of lexicographers who are working on the dictionary of the Royal Spanish Academy and which we believe to be around 20 people. If this team worked with the described tool and method they would be able to produce about 180.000 senses per year. If they worked three years it would run into more than half a million senses and after only five and a half year the number would reach one million senses, i.e. the most comprehensive Spanish dictionary ever produced. In this respect, the gauntlet is down! Basically, it is a question of fully adapting to and exploiting the new technologies and techniques put at the disposal of lexicography. The Online Dictionaries of Spanish Valladolid-UVa provides one example of how this can be done although we are not claiming that it is the only road that leads to Rome.

8. The lexicographer's competences and active role

Tools, methods and empirical bases do not make a dictionary on their own. However advanced the technology, the most important thing in dictionary making is still the human factor in the form of skilled, knowledgeable and motivated lexicographers. Dictionaries, as Gudmann (2014: 31) rightly states, "are still made by real human beings through a creative process without a correct answer carved in stone."

So what precisely is required from the lexicographers participating in the project? Here it is once more necessary to make a distinction between *knowledge* and *skills*. This means, on the one hand, that a modern online project as the Online Dictionaries of Spanish Valladolid-UVa cannot prosper without a manager (lexicographer-in-chief) who has a profound knowledge of lexicographical theory and methodology as well as the ability to design a dictionary concept, write instructions, select and train a team of collaborators, and supervise the daily work.

On the other hand, it also means that the project needs a team of skilled lexicographers who are highly productive and able to generate lexicographical data of the required quality. These practical lexicographers should above all have linguistic competences in Spanish, i.e. they should be native Spanish speakers. In addition, they should also have "a sound knowledge of the world generally and about at least one specific field" (Bergenholtz 2013: 5). The specific knowledge could be about linguistics but it could also be about other disciplines relevant to the project. In this respect, Bergenholtz (2013) reports that the team of lexicographers working on the Danish Internet Dictionaries is composed of people from language studies, mathematics, chemistry, molecular biology, physics, legal science, economics and chemistry. When the Spanish lexicographers were tested before being employed in the Valladolid-UVa project they should, apart from language competence and knowledge, also prove other relevant skills such as their ability to use computers, navigate on the Internet, find relevant data according to the instructions, and transform these data into easily understandable Spanish definitions.

However, this is only the starting point. In some of the actions listed in Section 5, it becomes clear that the selection of meaning items is not an exact science with "a correct answer carved in stone". On the contrary, a successful result depends to a large extent on the lexicographer's active role and decisions, which are not only based on his or her language competences and knowledge, but also on experience. This is, at least, the case for the following actions:

— Action 4: How many pages should be skipped?

— Action 6: Which parts of the minitexts are relevant?

— Action 8: How many pages should be reviewed?

— Action 9: When is the amount of empirical data satisfactory?

— Action 11: When is the lexicographer satisfied with the process?

It goes without saying that the decisions taken in these cases will affect the quality of the final product and that the decision time itself will have consequences for productivity. The challenge is therefore to reduce decision time and raise the quality of the decisions. In order to understand what happens, or should happen, it seems beneficial to refer to the "five stages of skill acquisition" proposed by Dreyfus and Dreyfus (1986) who operate with the following types of performer according to skills: novice, advanced beginner, competent performer, proficient performer and expert. Flyvbjerg (2001) has summarized the characteristics of these five stages in the learning process:

> (1) *Novices* act on the basis of context-independent elements and rules. (2) *Advanced beginners* also use situational elements, which they have learned to identify and interpret on the basis of their own experience from similar situations. (3) *Competent performers* are characterized by the involved choice of goals and plans as a basis for their actions. Goals and plans are used to structure and store masses of both context-dependent and context-independent information. (4) *Proficient performers* identify problems, goals, and plans intuitively from their own experientially based perspective. Intuitive choice is checked by analytical evaluation prior to action. (5) Finally, *experts'* behaviour is intuitive, holistic, and synchronic, understood in the way that a given situation releases a picture of problem, goal, plan,

> decision, and action in one instant and with no division into phases. This is the level of true human expertise. Experts are characterized by a flowing, effortless performance, unhindered by analytical deliberations. (Flyvbjerg 2001: 20-21)

Flyvbjerg (2001: 21) adds that the above model contains a "qualitative jump" from the three first stages to stage 4 and 5, and that "the jump implies an abandonment of rule-based thinking as the most important basis for action, and its replacement by context and intuition". In another publication, Dreyfus and Dreyfus (1992) emphasize this point:

> It seems that beginners make judgments using strict rules and features, but that with talent and a great deal of involved experience the beginner develops into an expert who sees intuitively what to do without applying rules and making judgments at all. The intellectualist tradition has given an accurate description of the beginner and of the expert facing an unfamiliar situation, but normally an expert does not deliberate. He does not reason. He does not even act deliberately. He simply spontaneously does what has normally worked and, naturally, it normally works. (Dreyfus and Dreyfus 1992: 117)

If this model is transferred to lexicography, it is easier to understand what makes a good lexicographer and what should be taken into account when selecting a team of collaborators for a dictionary project.

In the concrete case of the Valladolid-UVa project, when the job advert was posted, more than 90 candidates sent in their applications. Of these, 12 applicants were pre-selected based on their documented knowledge and competences. They were then (February 2014) offered a 30 hour course taught by Pedro A. Fuertes-Olivera and Helene R. Gudmann, the latter a skilled lexicographer with experience from the Danish Internet Dictionaries. The course included introduction to lexicography and the Valladolid-UVa project as well as instructions on how to collect data on the Internet, write definitions, and prepare the remaining data categories. The 12 candidates then took a test where they should fill in a number of cards in the database based on the instructions. The four best performers were selected for the job and started working for a 3-month trial period which is a requirement of the Spanish legislation.

When they took the test, the four lexicographers who eventually got the job, could most precisely be characterized as *novices* according to the Dreyfus model. In the 3-month trial period they were expected to develop relatively quickly into *advanced beginners* and then into *competent performers*, a transformation that would be reflected in growing productivity and quality of their lexicographical work. If this were not the case, the contract would be cancelled. However, after three months they will more often than not still be very much dependent on lexicographical instructions (rules) and chosen goals and plans as a basis for their actions and decisions, although some of them may start using their intuition based on previous experience. If this happens, they will jump to stage 4, that of the *proficient performer*, which is the minimum level that should be expected from all lexicographers participating in projects of the magnitude and importance of the Online Dictionaries of Spanish Valladolid-UVa.

Stage 5 in the Dreyfus model is the *expert* or *virtuoso* level which is only reached by a section of practicing lexicographers. It depends above all on experience but cannot be reached without talent which should be spotted in the trial period. At this stage, the lexicographer simply acts spontaneously without deliberating, and therefore "even the best lexicographers, when pressed, can never explain what they are doing, or why" (Wierzbicka 1985: 5). Lexicography has now become *pure art and craft*. However, this does certainly not imply that "lexicography has no theoretical foundation", as Wierzbicka also claims, or that the lexicographical compilation process is not based on rules. It rather signifies that these rules have been completely internalised and integrated with experience-based intuition into a flowing, effortless and holistic performance where the lexicographer, as any other person who performs at this level, cannot explain what he or she is doing. The lexicographical theory and instructions (rules) are still there, at the bottom of everything, as an important instrumentarium that is needed in order to transmit knowledge and skills to future lexicographers.

Today, the lexicographers participating in the project at the University of Valladolid can be characterized as either proficient performers or experts as defined in the Dreyfus model. This implies that they are now able to take quick, qualified and intuitive decisions to act in situations like the ones discussed above (Actions 4, 6, 8, 9 and 11) as well as in all other situations that may be related to the compilation process. In this respect, *experience-based human intuition is an important production factor without which success in a lexicographical project would be impossible*.

All this can be summarized as follows: Firstly, the technology and methods needed to work with the Internet directly as a corpus require skilled, knowledgeable and talented human beings who are motivated to make an extraordinary performance; and secondly, these characteristics should be spotted by the project manager as early as possible, an ability which also demands experience and talent. This is at least the experience from the on-going work on the Online Dictionaries of Spanish Valladolid-UVa.

9. Conclusions

The corpora were introduced in the 1960's whereas the Internet as a generalized phenomenon did not see the light of the day until three decades later, in the 1990's. It is surprising that lexicography so far has made more use of an old technology than of a more recent one. The question is whether it is time to explore the lexicographical possibilities of the Internet. The experience of the Dictionaries Valladolid-Uva clearly indicates that the time is more than ripe. It shows that skilled and well-trained lexicographers working with the right tools and methods are perfectly able to handle the undeniable disadvantages when using the Internet directly as a corpus instead of the traditional text corpora. This, of course, does not imply that these corpora are no longer of relevance to lexicography as they still have an important role to play when performing a number of tasks. It means above all that lexicography in order to confront its current crisis needs to go online not only to present its products but also to make them with the necessary quality and productivity.

In this paper we have discussed a set of online dictionaries of Spanish which is one of the world's biggest languages with more than four hundred million native speakers. It is evident that the number of Spanish texts placed on the Internet is enormous. However, as the experience of the Danish Internet Dictionaries shows, it is perfectly possible to use the same technology and methodology when working with a smaller language with only five million first-language speakers. In this respect, the possible problem is not as much the number of speakers as the penetration and generalized use of the Internet within a given speech community. This suggests that there may be some African languages with relatively few speakers where the collection of Internet-based texts is still not big enough to compile dictionaries as described in this contribution but they will be the exception to the rule. In most cases, the amount of Internet-based texts is already sufficient, or will be it in the nearby future. The Internet is here to stay, at least for a period of years, and it could easily be a big mistake not to start exploring its lexicographical possibilities already today.

Acknowledgements

Thanks are due to the Spanish Ministry of Economy and Competitiveness for funding the project "La Teoría Funcional de la Lexicografía: Diseño y Construcción de Diccionarios de Internet" (Ref. FFI2014-52462-P).

References

Dictionaries

Bergenholtz, H. (Ed.). 2016. *De Danske Netordbøger.* Odense: Ordbogen.com.

Fuertes-Olivera, P.A. and H. Bergenholtz (Eds.) in collaboration with M.Á. Sastre Ruano, E. Álvarez Ramos, M. Fonseca Herández, M.J. López Carrero, Á. Prieto Salvador and O. Saldaña. *Diccionarios en Línea de Español "Universidad de Valladolid".* Hamburg: Lemma.com. (Under construction.)

Moliner, M. 2007. *Diccionario de Uso del Español*. Third edition. Madrid: Gredos.

Real Academia Española. 2014. *Diccionario de la Lengua Española*. 23rd edition. Madrid: Espasa.

Sánchez Pérez, A. (Ed.). 2001. *Gran Diccionario de Uso del Español Actual***.** Madrid: Sociedad General Española de Librería.

Seco, M., O. Andrés and G. Ramos. 2011. *Diccionario del Español Actual*. Madrid: Aguilar.

Other Literature

Atkins, B.T.S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.

Bergenholtz, H. 1996. Korpusbaseret Leksikografi. *LexicoNordica* 3: 1-15.

Bergenholtz, H. 2013. The Role of Linguists in Planning and Making Dictionaries in Modern Information Society. Kwary, D., N. Wulan and L. Musyahda. (Eds.). 2013. *Lexicography and Dictionaries in the Information Age. Selected papers from the 8th ASIALEX International Conference*: 1-10. Surabaya: Airlangga University Press.

Bergenholtz, H. and H. Agerbo. 2014. Meaning Identification and Meaning Selection for General Language Monolingual Dictionaries. *Hermes* 52: 125-139.

Bergenholtz, H. and S. Tarp (Eds.). 1995. *Manual of Specialised Lexicography: The Preparation of Specialised Dictionaries.* Amsterdam/Philadelphia: John Benjamins.

Dreyfus, H. and S. Dreyfus. 1986. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York: Free Press.

Dreyfus, H. and S. Dreyfus. 1992. What is Moral Maturity? Towards a Phenomenology of Ethical Expertise. Ogilvy, J. (Ed.). 1992. *Revisioning Philosophy*: 111-131. Albany, NY: State University of New York Press.

Flyvbjerg, B. 2001. *Making Social Science Matter. Why Social Inquiry Fails and How It Can Succeed Again.* Cambridge: Cambridge University Press.

Francis, W.N. 1979. Problems of Assembling and Computerizing Large Corpora. Bergenholtz, H. and B. Schaeder (Eds.). 1979. *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora*: 110-123. Königstein/Ts.: Scriptor.

Fuertes-Olivera, P.A. 2012. Lexicography and the Internet as a (Re-)source. *Lexicographica* 28: 49-70.

Fuertes-Olivera, P.A. and H. Bergenholtz. 2015. Los Diccionarios en Línea de Español "Universidad de Valladolid". *Estudios de Lexicografía* 4: 71-98.

Fuertes-Olivera, P.A. and S. Tarp. 2014. *Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminography.* Berlin/Boston: De Gruyter.

Gouws, R.H. and S. Tarp. 2016. Information Overload and Data Overload in Lexicography. *International Journal of Lexicography* 29(4). (In print.)

Gudmann, H.R. 2014. *Betydningshuller i Spanske Ordbøger. En Undersøgelse af Betydningsenheder i Spanske Monolingvale Almene Receptionsordbøger.* M.A. Thesis. Aarhus: Aarhus University, Department of Business Communication.

Gudmann, H.R. 2015. Lagunas de Significado en los Diccionarios Españoles. *Estudios de Lexicografía* 4: 161-184.

Hanks, P. 2012. The Corpus Revolution in Lexicography. *International Journal of Lexicography* 25(4): 398-436.

Kilgarriff, A. 1997. I Don't Believe in Word Senses. *Computers and the Humanities* 2(31): 91-113.

Kilgarriff, A. 2012. [Review of] Pedro A. Fuertes-Olivera and Henning Bergenholtz (Eds.). 2012. *e-Lexicography: The Internet, Digital Initiatives and Lexicography. Kernerman Dictionary News*, July 2012: 26-29.

Kilgarriff, A. and G. Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29(3): 333-347.

Lees, R. 1962. Oral contribution. Quoted by Francis, W.N., 1979: 110.

Rundell, M. 2015. From Print to Digital: Implications for Dictionary Policy and Lexicographic Conventions. *Lexikos* 25: 301-322.

Sinclair, J.M. 1997. Introduction. *Collins Cobuild English Language Dictionary*: xv-xxi. London: HarperCollins.

Tarp, S. 2008. *Lexicography in the Borderland between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography.* Tübingen: Max Niemeyer.

Tarp, S. 2015. Structures in the Communication between Lexicographer and Programmer: Database and Interface. *Lexicographica* 31: 217-246.

Tarp, S. 2016. Excesos en el Uso de Corpus en la Lexicografía: «Pesca» de Términos y Definiciones. *Revista de Lexicografía* 21. (In print.)

Wierzbicka, A. 1985. *Lexicography and Conceptual Analysis*. Ann Arbor: Karoma.

Xue, M. and S. Tarp. 2016. Corpus-based, Corpus-driven or Corpus-assisted lexicography? The Limited Usefulness of Corpora in Defining Specialised Terms. *Lexicographical Studies* 4: 1-11.