EDICIONES COMPLUTENSE

Círculo de Lingüística Aplicada a la Comunicación
ISSN-e: 1576-4737

ARTÍCULOS

Innovations in Spanish Lexicography: The *Diccionario Digital del Español (DIDES)*

Pedro A. Fuertes-Olivera
International Centre for Lexicography (Universidad de Valladolid, Spain)
Department of Afrikaans and Dutch (University of Stellenbosch. South Africa)
Centre of Excellence in Language Technology, Ordbogen A/S, Odense, Denmark

https://dx.doi.org/10.5209/clac.90467

Abstract: In 2014, the International Centre for Lexicography, a research group at Valladolid signed a contract with Ordbogen A/S (a Danish language technology company) and the University of Valladolid for developing a lexicographic project, the so-called *Diccionarios Valladolid-UVa* (Fuertes-Olivera 2019 and 2022; Fuertes-Olivera & Tarp, 2022; Fuertes-Olivera et al, 2018; Tarp and Fuertes-Olivera, 2016). Each partner gave around 180,000 euros (the International Centre for Lexicography's contribution came from the research projects FFI2011-22885, VA067A12-1 and FFI2014-52462-P), to be employed in the design and construction of Spanish dictionaries (in particular, a general dictionary of Spanish, a Spanish dictionary of accounting, a bilingual Spanish-English/English-Spanish dictionary and a bilingual Spanish-English/English-Spanish accounting dictionary). By the end of 2019, the Spanish Research Agency and the regional one had stopped funding the research projects of the International Centre for Lexicography, which prompted a change of course in the project and the actual dismemberment of the research group. At that time, the project had completed around 50,000 dictionary articles and was cooperating with Ordbogen in launching in Spain *Write Assistant*, an integrated writing and translation tool (Fuertes-Olivera and Tarp, 2020). In July 2020, the editor of the project decided to continue working on it, concentrating on three tasks. Firstly, he was going to create more dictionary articles for the general dictionary of Spanish (i.e., the accounting and bilingual projects were abandoned). Secondly, he would accommodate the existing dictionary articles to the new ones to new findings (Tarp, 2022). And finally, both he and Sven Tarp would explore a way ahead for making the lexicographic data truly innovative and developing more effective products, services, processes, technologies, and business models. This article explains decisions concerned with the *Diccionario Digital del Español (DIDES*, the name given to the general dictionary of Spanish of the abovementioned lexicography project is accessed at https://diesgital.com). This online dictionary was released by Ordbogen in June 2023, and some of its innovations are concerned with the following: (a) selection of lemmas; (b) social mores; (c) the use of the Internet as a lexicographic source; (d) the treatment of lexicographic data; (e) the use of technology for searching, presenting, and updating the dictionary articles; and (f) the introduction of new business models. Since August 2023, I am also using generative Artificial Intelligence for crafting definitions and completing the lexicographic data of each dictionary article. For reasons of space, I have decided to explain how this innovation works in future publications (Fuertes-Olivera, 2025).
Keywords: lexicography; online dictionary of Spanish; innovations; DIDES; lemma list; log files; social mores; lexicographic sources; lexicographic technology; data presentation; business models.

How to cite: Fuertes-Olivera, P. (2025). Innovations in Spanish Lexicography: The *Diccionario Digital del Español (DIDES). Círculo de Lingüística Aplicada a la Comunicación* 103 (2025): 199-213. https://dx.doi.org/10.5209/clac.90467

1. The Lexicographic Project *Diccionarios Valladolid-UVa*

The lexicographic project *Diccionarios Valladolid-UVa* started officially in January 2014, with the signing of a contract between the Danish company Ordbogen A/S, the University of Valladolid, and the International Centre for Lexicography research group, each committing 180,000€ to the project; this would be spent in the next four or five years. In the same month, we selected 4 part-time lexicographers, each with a 19-hour-a-week work schedule and with an annual cost of around 25,000 € (salary + labor expenses) per lexicographer. The selection process consisted of two stages, the first of which was devoted to examining the CV and English proficiency of 50 applicants. This stage resulted in the shortlisting of 10 applicants, who were given a 30-hour crash course on how to write dictionary articles and search lexicographic data with Google. These ten applicants were then asked to write 10 dictionary articles, which had been selected by the editor of the project, in a controlled environment. Their answers were then evaluated by three researchers of the International Centre for Lexicography, who selected 4 of the 10 applicants. These four lexicographers started their work in March 2014; they all worked for 4 hours from Monday to Thursday and 3 hours on Friday. They were in the same room, next to the office of the editor of the project, who could check and answer their doubts very easily and quickly. They worked on the project until June 2020, (the Spanish Research Agency and the regional research authorities had decided to stop funding the research projects they had been financing up to that time (The reference numbers of the financed projects are as follows: FFI2011-22885, VA067A12-1 and FFI2014-52462-P).

The project was based on three main ideas (see Fuertes-Olivera, 2019, for a more detailed analysis). Firstly, it argued that dictionaries are reference tools conceived for consultation with the genuine purpose of meeting specific information needs experienced by specific types of potential users in specific types of extra-lexicographical contexts (Bergenholtz & Tarp, 2003; Tarp, 2008; Fuertes-Olivera & Tarp, 2014). Dictionaries must be designed to assist their users by providing manual or automatic access to lexicographic data, either prepared by lexicographic teams or recommended by the team and extracted from open linked data, e.g., figures and Wikipedia links.

Secondly, it also claimed that, as they are reference tools, dictionaries must be prepared, designed, and compiled as up-market products and/or services, i.e., tools that displace established competitors by making use of disruptive technologies. For instance, preparing *dynamic dictionary articles*, i.e., different data for different users in different situations is a feature of upmarket online dictionaries that can easily be implemented as a lexicographic strategy for broadening the customer base of online dictionaries.

Finally, it assumed that we are in the middle of a data-driven economy, and consequently lexicographers should prepare lexicographic data for coping with the following: *pervasive information asymmetry*, i.e., users should have at their disposal many information channels and will use the one(s) more useful for them; the *industrialization of learning through artificial intelligence*, e.g., the use of machine learning and neural networks for developing assistants and other auxiliary tools; and *new lexicographic uses*, e.g.. for discovering which words users search for (Fuertes-Olivera & Tarp, 2020; Tarp, 2022).

Cancelling public funding for the International Centre for Lexicography forced the project to change course. Since mid-2020, only the editor of the project has been engaged in it on a regular basis. He is totally committed to creating more dictionary articles for the general dictionary of Spanish, to adapting the existing dictionary articles to new findings (Tarp, 2022) and, together with Sven Tarp, to explaining the decisions taken; it is assumed that these are truly innovative, i.e., they are the result of the development of more effective products, services, processes, technologies, and business models.

This article assumes that lexicography cannot be achieved without innovation, and thus an explanation of this is given in the contexts of lemmas (section 3), social mores (section 4), lexicographic sources (Section 5), the treatment and presentation of lexicographic data (Section 6), technology (Section 7) and business models (Section 8). The innovations are illustrated with examples taken from the *Diccionario Digital del Español (DIDES)* (section 2). A conclusion will summarize the main ideas discussed.

2. The *Diccionario Digital del Español (DIDES)*

The *Diccionario Digital del Español (DIDES)* is the name given to the general dictionary of Spanish designed and compiled within the framework of the lexicographic project "*Diccionarios Valladolid-UVa*". The dictionary was released in June 2023. Figure 1 shows a screenshot of the data stored so far.

Here are some of the numerical details:

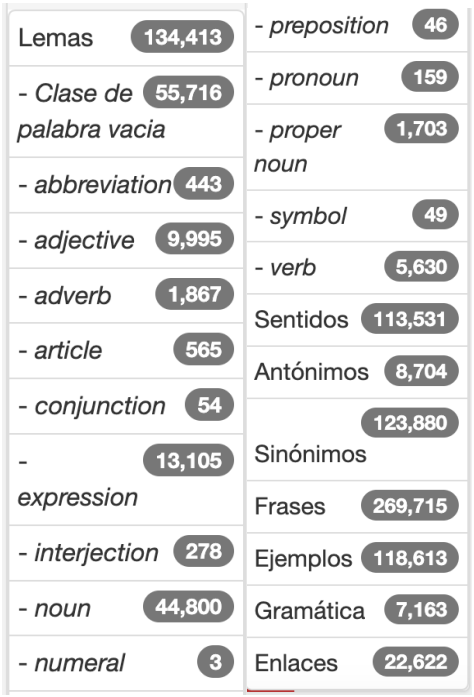
- It has more than 134,000 lemmas (section 3, below).
- Around 78,000 lemmas contain some lexicographic data, whereas around 55,000 are totally empty ("Clase de palabra vacía).
- Around 50,000 of the 78,000 lemmas examined have now been completed.
- There are more than 13,000 expressions, i.e. lemmas composed of 3 or more single words.
- Almost 30% of the lemma list in the dictionary are nouns; this figure is relevant and shows the key role nouns play in any language.
- There are more than 118,000 examples and 269,000 "frases" (i.e., chunks of texts that show some relevant information about the lemma). They have basically the same information, and hence will be placed under the same heading in the dictionary article (section 6).
- There are more than 113,000 meanings; these correspond to the completed lemmas (see section 3, below).

– There are more than 123,000 synonyms and 8,000 antonyms referring to the abovementioned meanings; these will offer several options and will allow the creation of "semantic and functional patterns", i.e., synonyms and/or antonyms for disambiguating meaning at a quick glance (section 6, below).
– There are more than 22,000 links, most of them to figures, Wikipedia articles and some YouTube videos and clips (section 6, below).
– There are more than 7,000 grammar notes, i.e., information on some relevant grammatical information (see section 6, below).

The above data indicate that *DIDES* is an on-going, large lexicographic project, whose main innovations are analyzed in the following sections. It is hoped that these will result in a sustainable lexicographic project.

Sustainability in lexicography does not refer to the resource (language) but to the financial resources that are needed for designing, making, and maintaining any lexicographic project (Colman, 2016). The sections below, all of which refer to innovations in Spanish lexicography, assume that these innovations are totally necessary for convincing funders and the Spanish-speaking world that they need more than "copycats" and "faster horses" (Tarp, 2011: 58-60) to meet their information needs. While it is true that people may just "google" what they need in many situations and that their needs are often satisfied, I think that this can be improved by offering them the possibility of consulting high-quality dictionaries such as *DIDES*.

Figure 1. Statistics of the *Diccionario Digital del Español.*



Source: Editor of the Dictionary (see section 5, below)

3. Innovations Connected with Lemmas

Innovation is "the practical implementation of ideas that result in the introduction of new goods or services or improvement in offering goods or services." (Innovation: Wikipedia) This definition indicates that innovation often takes place through the development of more effective products, processes, services, technologies, business models and so on, and that innovation is related to, but different from, invention. Hence, lexicographers do not have to invent the lexicographic wheel when they work on a new lexicographic project. They can and must use existing lexicographic resources, although in a novel and enhanced way; for example, with the lemma lists of existing dictionaries.

Fuertes-Olivera (2022) considers the selection of the headword or lemma list to be an *ongoing process*, i.e., one that is never finished. As such, lexicographers must decide on the method for selecting the initial lemma list and continuously enlarging it. Since the advent of the *Cobuild Dictionary* (Sinclair, 1987), lexicographers have mostly defended a corpus-based approach to headword selection, i.e., the words to be included must be *basically* extracted from corpora in accordance with their frequency and/or importance. My proposal is different: selection is a process that should take into consideration its inception and continuous development. Its initial stage aims at selecting the words that users *really* look up, as research has discovered that many of the words lemmatized in existing dictionaries — some researchers claim almost 80%; see Bergenholtz and Norddahl, 2014 — have never been looked up (Trap-Jensen et al. 2014). The *Diccionarios Valladolid-UVa* have followed this methodology and have initially selected two lists of single-word lemmas,

one for English and one for Spanish. The initial headword lists of the *Diccionarios Valladolid-UVa* were selected at the Ordbogen A/S headquarters, the Danish language technology company with which we have been designing and carrying out our lexicographical projects since 2014.

The Danish company used big data analytics for around two months. The process comprised several stages and was based on an analysis of around one million daily searches in 10 dictionaries, e.g., an English–Spanish/Spanish–English dictionary, an English–German/German–English dictionary, an English monolingual dictionary, a Spanish monolingual dictionary, and so on. It was possible to match around 80% of the searches, i.e., they were found in, at least, 8 of the log files of the ten dictionaries. These were considered ideal candidates for the initial lemma list in English and Spanish, each comprising around 20,000 words, 16,678 of which were used for starting *DIDES*. An analysis of the process and the list highlights the main characteristics of the initial lemma list:

- Users search for words of very low frequency in reference corpora such as CREA and CORPES XXI, even for words that are not included at all in reference corpora. For example, the words *Balbino* (a Roman emperor murdered by pretorians), *madison* (a type of dance and a kind of cycling competition), *mae* (an informal means of address for addressing young people used in South America), *sobrehipotecar* (a technical term used in law and economics referring to the illegal process of taking more debt than the value of the property mortgaged) and *ostomía* (a technical term in medicine referring to a type of surgery that allows bodily waste to pass through a surgically created stoma on the abdomen into a prosthetic known as "ostomy bag), are in the log files but not in the lemma list of Spanish dictionaries and have very low frequencies in corpora (e.g., *sobrehipotecar* has zero concordances in the above-mentioned reference corpora, perhaps because this word was introduced in 2008 by the European Central Bank in connection with the chaos resulting from the bankruptcy of Lehman Brothers).
- Users search for words connected with their daily lives, typically health conditions, organizations, plants, animals, and tools. For instance, words such as *OCDE, colectivo LGTBI, cachí* (a bird) and *out* (as a noun, adjective and interjection) are not lemmatized in *Diccionario de la Lengua Española* (DLE), although they are frequently used in the Spanish- speaking world.
- Users also search for words that are mostly or only used in America (i.e., they are "americanismos"). This clearly indicates the necessity of paying attention to these words. For example, words and expressions such as *abombe, bacho, buen pago, cablevision, cachí*, and so on, are in the log files but are not in the lemma list of DLE.

The initial lemma list must be systematized and permit enlargement, i.e., the process used for adding more lemmas to the initial lemma list. Systematization means that all the members of the lists must be converted into a *unit of inclusion*, e.g., a lemma in traditional lexicography. Following standard practice, the editor *initially* converted the list into 16,678 single-word lemmas and these were included in the Dictionary Writing System (DWS) or the editor of the *DIDES* in their canonical form, e.g., the infinitive of the verb, but adapted to an online search process (section 7, below).

Enlargement is also an *on-going process*. It is initially concerned with the words and expressions that are related with the lemmas of the initial lemma list. In the *DIDES*, I have taken the following decisions, which are innovative in Spanish lexicography:

- I have eliminated all constructs such as *professor, ra*, *quieto, ta* and so on. In *DIDES*, all lemmas are real words and expressions, e.g., *profesor* and *profesora* are two lemmas (section 4).
- Homonyms are distinguished according to their word class, inflection(s), if any, and the articles with which they agree. This means that *agudo* is lemmatized twice (*agudo* as a noun goes with *un agudo, el agudo, unos agudos, los agudos* and *agudo* as an adjective goes with *agudo, aguda, agudos* and *agudas*). In a similar vein, *policía* is lemmatized twice (*un policía, el policía, unos policías, los policías* and *una policía, la policía, unas policías* and *las policías*). Furthermore, *casa* is also lemmatized twice (*una casa, la casa, unas casas, las casas* and *casa*, without any inflection or morphological change when it is used figuratively to refer to an imaginary place where a person or organization is or feels safe, as in the example "en este lugar me siento en casa") (section 6).
- I have lemmatized all related words, i.e., those that stem from the initial single-word lemmas due to grammar rules. In Spanish, these *basically* affect some nouns, adjectives, adverbs, and verbs. For instance, *abanderado* is a male noun and its related word is *abanderada* (female noun). In traditional Spanish dictionaries such as the DLE, this process of enlarging only exists for lemmatizing some manner adverbs, i.e., they are formed by adding *-mente* to the base of an adjective, e.g., *abiertamente*. For the rest of related words, Spanish dictionaries use constructs such as *abanderado, ra* that do not exist in real linguistic interactions (Fuertes-Olivera and Tarp 2022) or they do not lemmatize them at all. For instance, the related word of the verb *peinar* is *peinarse*, and the related word of the adjective *abierto* is the noun *abierto*. These and derived nouns (*abastos*, plural nouns), informal adverbs and interjections (*claro*) are lemmatized in *DIDES*.
- I have lemmatized all expressions found during the compilation process of the dictionary articles covering the initial lemma list (section 6, below). An expression or "extended unit of meaning" (Rundell 2018) is a linguistic unit formed by three or more orthographical words that expresses a concept and is used as a unit within a sentence. Such a unit is converted into an "extended-unit-of-meaning lemma" and included in the lemma list if it is still in use, e.g. by being in approximately 5% of the Google minitexts

used as sources (section 5) and in four out of seven existing dictionaries that I also referenced during the process of compilation: *Diccionario de la Lengua Española* (DLE); *Diccionario del Español Actual* (Seco et al. 2011); *Diccionario Español–Inglés* (Collins); *Diccionarios.com; Diccionario español de Google* (Google); *SpanishDict*; and *WordReference (Spanish; Spanish–English)*. Fuertes-Olivera (2022) claims that the lemmatization of expressions is based on the tenets of *semantic network theory* (Forster & Chambers, 1973). This theory affirms that humans *mostly* use meaning networks in their daily linguistic interactions. For instance, the Spanish adjective *agudo* has 14 different meanings or senses in the DLE and 5 expressions that are included as run-ons (*acento agudo; ángulo agudo; octava aguda; octavilla aguda*; and *poliomelitis aguda*). In *DIDES, agudo* has 13 meanings as an adjective and is part of 12 more lemmas: the five run-ons of the DLE and a further 7 not in this dictionary: *tono agudo*, *verso agudo, zumbido agudo*, *silbido agudo*, *lumbago agudo*, *abdomen agudo* and *lino silvestre agudo*. This process is a very active one, and I think that in two years' time, DIDES will have more than 30,000 "extended-unit-of-meaning lemmas", i.e., lemmatized expressions such as *quiosco de bebidas, comida de plástico, alojamiento y comida, beber la sangre, beber a gallete, beber los vientos por*, and so on.

To sum up, the process of lemmatization used in *DIDES* highlights six innovations. Firstly, the initial lemma list comes from log files, i.e., real searches, and not from corpora, literary works, or existing dictionaries. Secondly, as an on-going process, the lemmatization of new "realities" (linguistic and social) needs both the desire and the technology which allow lexicographers to incorporate them as soon as they are encountered (section 7). Thirdly, the initial lemma list is amplified by applying grammar rules, social mores, and "better search and find" technologies during the process of compilation of the dictionary articles. Fourthly, all the lemmas refer to existing linguistic and/ or social entities. Fifthly, it never uses run-ons, most of them being lemmas. Finally, homonyms are differentiated in terms of their word class, inflections, and the articles with which they agree. The rationale for such a philosophy is twofold: (a) it offers a better description of the language and (b) it facilitates searching and retrieving. As a consequence, the dictionary might be better prepared for using NLP tools. Table 1 shows 8 lemmas of DIDES and their treatment in *Diccionario de la Lengua Española* (DLE); *Diccionario del Español Actual* (Seco et al. 2011); *Diccionario Español de Google* (Google); and *WordReference (Español)*. Neither of them is lemmatized in DEL, Seco et. al., Google, and Word Reference.

Table 1. Lemmas of DIDES and their lexicographic treatment in selected dictionaries

Lemmas in DIDES	DLE	Seco et al.	Google	Word Reference
principio de autonomía	Not found	Not found	Not found	Not found
administración de loterías	Not found	Not found	Run-on in *administración*	Run-on in *administración*
profesora	Not found. It forces users to deduce that it is part of the lexicographic construct *profesor, ra*	Not found. It forces users to deduce that it is part of the lexicographic construct *profesor, -ra*	Found in the construct *profesor, profesora*	Not found. It forces users to deduce that it is part of the lexicographic construct *profesor, ra*
peinarse	Not found. It forces users to deduce that one of the meanings of *peinar could be peinarse* by understanding the formula "U.t.c. prnl."	Not found. It forces users to deduce that one of the meanings of *peinar could be peinarse* by reading the usage note "*Frec el cd es reflexivo*" (the complement is often reflexive)	Not found	Not found
estupendo (adverb)	Not found. It forces users to deduce that one of the meanings is an adverb by understanding the formula "U.t.c. adv."	Included as *adv.* in the lexicographic construct *estupendo -da*	included as *adverbio* in the lexicographic construct *estupendo, estupenda*	Not found
casa (without any inflection)	Not found	Not found	Not found	Included as a meaning of *casa* without any indication of its grammar and function.

alto (noun)	Not found. It forces users to deduce it by interpreting the formula "U.t.c.s." in *alto, a*	Not found. It forces users to deduce it by interpreting the formula "Tb n m." in *alto,ta*.	Indication of nombre masculino in the lexicographic construct *alto, alta*	abbreviations "m." and "f." in several meanings of the lexicographic construct *alto, ta*
comer a dos carrillos	a run-on in the lemma *carrillo*	a run-on in the lemma *carrillo*	link in the lemma *comer* without any information on the meaning and use of the expression	a run-on in the lemma *carrillo*

4. Innovations connected with Social Mores

Dictionaries are powerful ideological tools and have always been used for promoting (even for imposing) a specific representation of reality within a given context. For instance, the feminist movement in the English-speaking world has contributed to the creation of lemmas such as *chair, chairperson, police officer*, and so on, that aim at eliminating the gender bias described by many scholars (Nissen, 1986; Fuertes Olivera, 1992; Holmes and Meyerhoff, 2003). Fuertes-Olivera and Tarp (2022) have also proposed several innovations aiming at eliminating the gender bias in general dictionaries of Spanish. These are included in *DIDES*:

- *DIDES* has two different lemmas for human beings, one referring to a man and another one to a woman. *DIDES* does not have lexicographic constructs such as *profesor, ra, maestro, tra, médico, ca*, and so on. In *DIDES*, there are two different lemmas: *profesor* and *profesora*, *maestro* and *maestra*, and *médico* and *médica* (section 3).
- Each of the above lemmas has one specific meaning, referring specifically to a man (*profesor*) or a woman (*profesora*), and one generic meaning referring to a person.
- *DIDES* prefers the specific lemmas to the generic ones. For instance, it has the lemma *fiscala* with two meanings and the following lexicographic note: "La forma "fiscala" favorece la visibilidad de la mujer en los cargos públicos" (the form "fiscala facilitates the visibility of women in the public sphere).
- *DIDES* does not usually include the meaning "wife of a professional man", which is sometimes recorded in Spanish dictionaries, as we consider it to be obsolete and out of touch with current social mores.

5. Innovations connected with lexicographic sources

Lexicographic data come from *lexicographic corpora*, defined by Fuertes-Olivera (2012: 51) as "any collection of texts where lexicographers can find inspiration for completing the dictionary structures they need when they are making a dictionary" and from any other source that can be used for the same purpose. The lexicographic sources of Spanish general dictionaries tend to be existing dictionaries, literary works, and corpora. *DIDES* is also different, as its main lexicographic source is the internet. Around 95% of all the lexicographic data used in the dictionary articles of *DIDES* are extracted from the internet. The intention is for the meaning and usage of any word or expression to be understood. Consequently, *DIDES* relies on "Google minitexts", i.e., the two to three lines Google retrieves when making a particular search, for an initial analysis of the meaning and usage of lemmas and homepage. Since 2023, it also uses generative AI chatbots as lexicographic sources (Fuertes-Olivera, 2025).

Tarp & Fuertes-Olivera (2016: 280-281) summarized the process of using Google minitexts as the main lexicographic source of *DIDES*. This process is now somewhat simpler, as the editor-in-chief of the project is now the only lexicographer working actively on the project:

- A lemma contained in the editor (i.e., the lexicographic database) is chosen and "googled" in inverted commas (section 6).
- A "traditional" Google-search result appears (see Figure 2).
- The first three pages are ignored because they typically contain existing dictionary articles and publicity.
- The minitexts appearing on each page are read to get a general idea of the subject matter.
- Using the "copy and paste" method, the relevant parts of the minitexts are copied onto a Word document.
- Simultaneously, examples, chunks of texts, synonyms, antonyms, and word formations (these are typically idiomatic expressions and multi-word lemmas; see section 3) are selected for incorporation in the respective fields of the editor (section 6).
- Several Google pages are reviewed until no further new data appear and everything is repeated. For multi-word lemmas, this process is quicker and easier than for single-word lemmas. Multi-word lemmas (i.e., expresiones in our lemma list) tend to have one or two meanings, one of them usually figurative (section 6).
- Once a satisfactory amount of empirical data has been selected, it is grouped according to meaning.
- Based on the data grouping, the first definitions are written according to new findings (section 6).
- At this stage the lexicographer decides if they are 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.

– Once the lexicographer has completed meaning selection and written the definitions concerning the lemma, the data are subjected to two additional processes. Firstly, the data found in the Google minitexts are compared with information existing in the following reference sources: *Diccionario de la Lengua Española* (DLE); *Diccionario del Español Actual* (Seco et al. 2011); *Diccionario Español–Inglés* (Collins); *Diccionarios.com; Diccionario español de Google* (Google); *SpanishDict*; and *WordReference (Spanish; Spanish–English)*. Any difference among them is checked; for example, to compare if a meaning described in, say, DLE still is in use. The checking takes place by performing "guided searches", which consist of googling the lemma between inverted commas and adding some features of the meaning. For example, "comer" + equipo rival + deporte for the figurative meaning "a sportsperson or team easily defeated another competitor". Secondly, the lemma is googled with the formula "Wikipedia" + "desambiguación" (section 6). This is important, as the analysis of the log files show that many of them are connected with health problems, plants, animals, tools, and processes, i.e., the terms Wikipedia typically describes. This search provides many new meanings of the lemma, most of which are absent in existing dictionaries of Spanish (section 6).

Figure 2 shows Google minitexts of the search "comer a dos carrillos", a colloquial expression that has a literal meaning (someone eats quickly and happily), and two figurative meanings (someone wants to have several, even competing, responsibilities at the same time; and something merits praise because of its high quality). The figurative meanings are not typically included in Spanish dictionaries. If necessary, I click on the homepage to check what the minitext indicates.

Figure 2. Results when googling "comer a dos carrillos" (excerpts).

6. Innovation connected with the Treatment of Lexicographic Data

Spanish lexicographers typically describe their lexicographic data by making recursive definitions, copying and pasting most of the data (especially definitions), and assuming that users are linguists who know the meaning and function of linguistic metadata (examples 1 to 4):

enseñante
1 adj. Que enseña. U.t.c.s

Example 1. The dictionary article **enseñante** in DLE

enseñante
adj [Pers.] que enseña [5]. *Frec n.* || *B Congreso* 28.11.80[In]. Este gasto ha supuesto disponer de ... 200.000 fichas informativas destinadas a personal enseñante. Diego *ABC* 21.8.63, 3: Él, con su otra legítima vocación de enseñante, de comunicante a los adolescentes de lo poco que ha aprendido.

Example 2. The dictionary article **enseñante** in Seco et al.

enseñante
adjetivo
1. Que enseña
2. *Nombre común*
Persona que ejerce la docencia en cualquiera de los niveles de instrucción en que se halla dividida la educación de un país o estado.
"pocas cosas alegran tanto a un enseñante como saber que sus palabras han despertado en otros el interés por aprender"

Example 3. The dictionary article **enseñante** in the Spanish dictionary of Google.

enseñante
1. *com.* Persona que se dedica a la enseñanza.

Example 4. The dictionary article **enseñante** in *Word Reference.* Español: definición

Examples (1) to (4) show the main characteristics of existing dictionaries of Spanish:

- They use recursive definitions, especially when they are the same (see adjective).
- They continue using abbreviations, e.g., DLE uses "U.t.c.s." for indicating that it can be nominalized.
- They do not include nor describe all possible categories and functions, e.g., *Word Reference* does not include the adjective function.
- They tend to limit the quantity of lexicographic data to the bare lemma (without inflections, conjugations, etc.), word class, definitions and, on some occasions, several examples. For instance, they do not include, inflections (e.g., the plural form), figures, links, and so on.
- They assume that users know the meanings or functions of linguistic metadata such as "com" in *Word Reference.*
- Their definitions are generally useless. For instance, only definition 2 of Google informs a potential user of the meaning of **enseñante**.

In other words, most of the lexicographic data of these four dictionaries is totally useless for most potential users. However, *DIDES* deals with lexicographic data on the basis of five innovations:

1. It never uses recursive definitions nor formulas such as "acción y efecto de escribir" for defining, say, "escritura" (writing) or deverbal nouns such as *creación, cribado, cruce,* and so on.
2. It never uses abbreviations or any (linguistic) formula that force users to convert the data into information. For example, some of the meanings of the verb "comer" go with the lexicographic note: "Verbo transitivo, es decir, va con un complemento directo" (transitive verb) or "Verbo intransitivo, es decir, va sin un complemento directo" (intransitive verb), depending on its function.
3. It offers different types of data, typically "words", "figures" (e.g., the definitions of material elements go with a photograph (see Figure 3, below), "videos" and links to external sources, e.g., to Wikipedia.
4. All definitions may possess the following accompanying data: examples and chunks of texts; synonyms and/or antonyms; geographical variations, if needed; diastratic variation, if needed; links to figures and/or external texts; different types of lexicographic notes, typically for grammar and/or usage; links to their own synonyms and/or antonyms, as well as examples and chunks of texts that illustrate the meaning and function defined.
5. All lemmas go with their inflections, articles, some conjugated forms (verbs) and grammar information, if needed. Examples (5-7) show these innovations:

enseñante (un enseñante, el enseñante, unos enseñantes, los enseñantes)
nombre
1. Definición
hombre que ejerce la docencia y da clases en cualquier nivel en que se halle dividido el sistema educativo de un país, región, ciudad, etc.
- **sinónimo:** *docente*
- **Frases y Ejemplos:**

1. y Juan es un enseñante en Málaga
2. Un enseñante de unos 40 años ha acudido a explicar el problema de matemáticas.

2. Definición
persona que ejerce la docencia y da clase en cualquier nivel en que se halle dividido el sistema educativo de un país, región, ciudad, etc.
- **sinónimo:** *docente*
- **Frases y Ejemplos:**

1. y estudiaban para ser los mejores enseñantes del país
2. Los enseñantes forman parte de un colectivo con mucho prestigio social en Finlandia

Example (5): The dictionary article **un enseñante, el enseñante, unos enseñantes, los enseñantes in** DIDES

enseñante (una enseñante, la enseñante, unas enseñantes, las enseñantes)
nombre
1. Definición
mujer que ejerce la docencia y da clases en cualquier nivel en que se halle dividido el sistema educativo de un país, región, ciudad, etc.
- **sinónimo:** *docente*
- **Frases y Ejemplos:**

1. y se casó con una enseñante, llamada Julia
2. La enseñante acudió a la reunión del sindicato acompañada por varios colegas.

Example (6): The dictionary article **una enseñante, la enseñante, unas enseñantes, las enseñantes** in DIDES

enseñante (enseñante, enseñantes)
adjetivo

1. **Definición**

Referido a una persona que ejerce la docencia y da clases en cualquier nivel en que se halle dividido el sistema educativo de un país, región, ciudad, etc.

- **sinónimo:** *docente*
- **Frases y Ejemplos:**

1 e iban cuatro personas enseñantes en el coche
2. A las personas enseñantes les gusta mucho estudiar para saber más

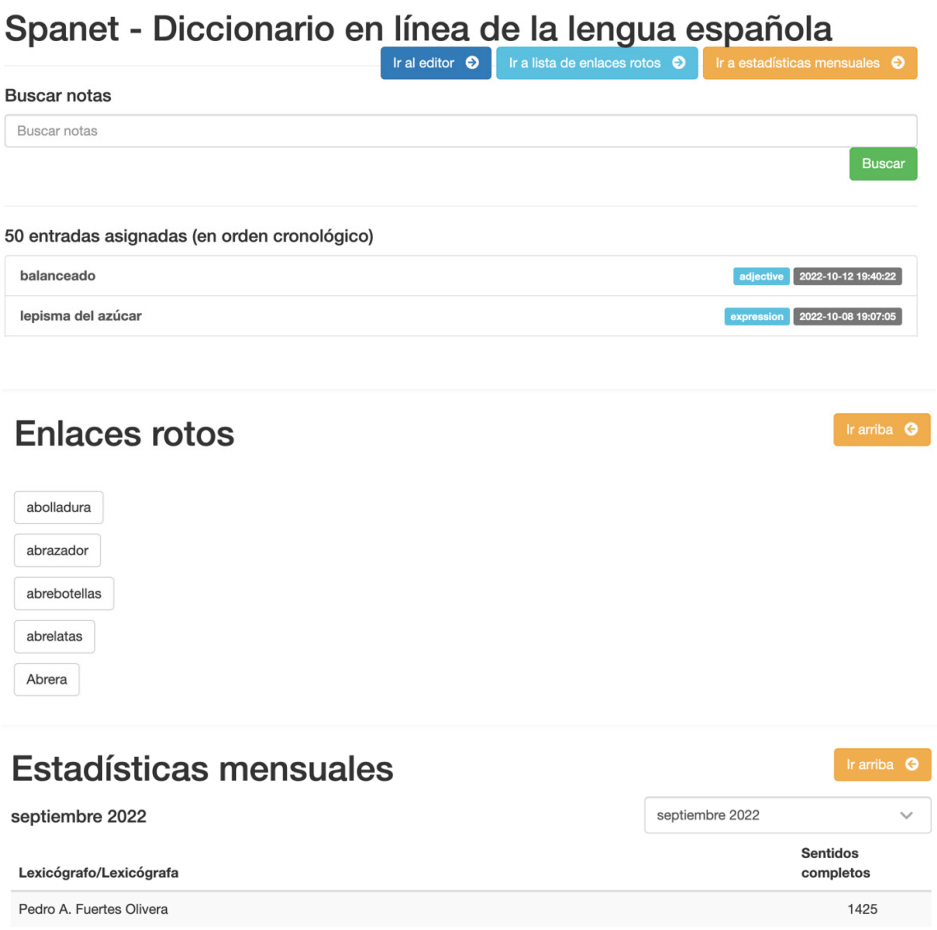
Example (7): The dictionary article **enseñante, enseñantes** in DIDES

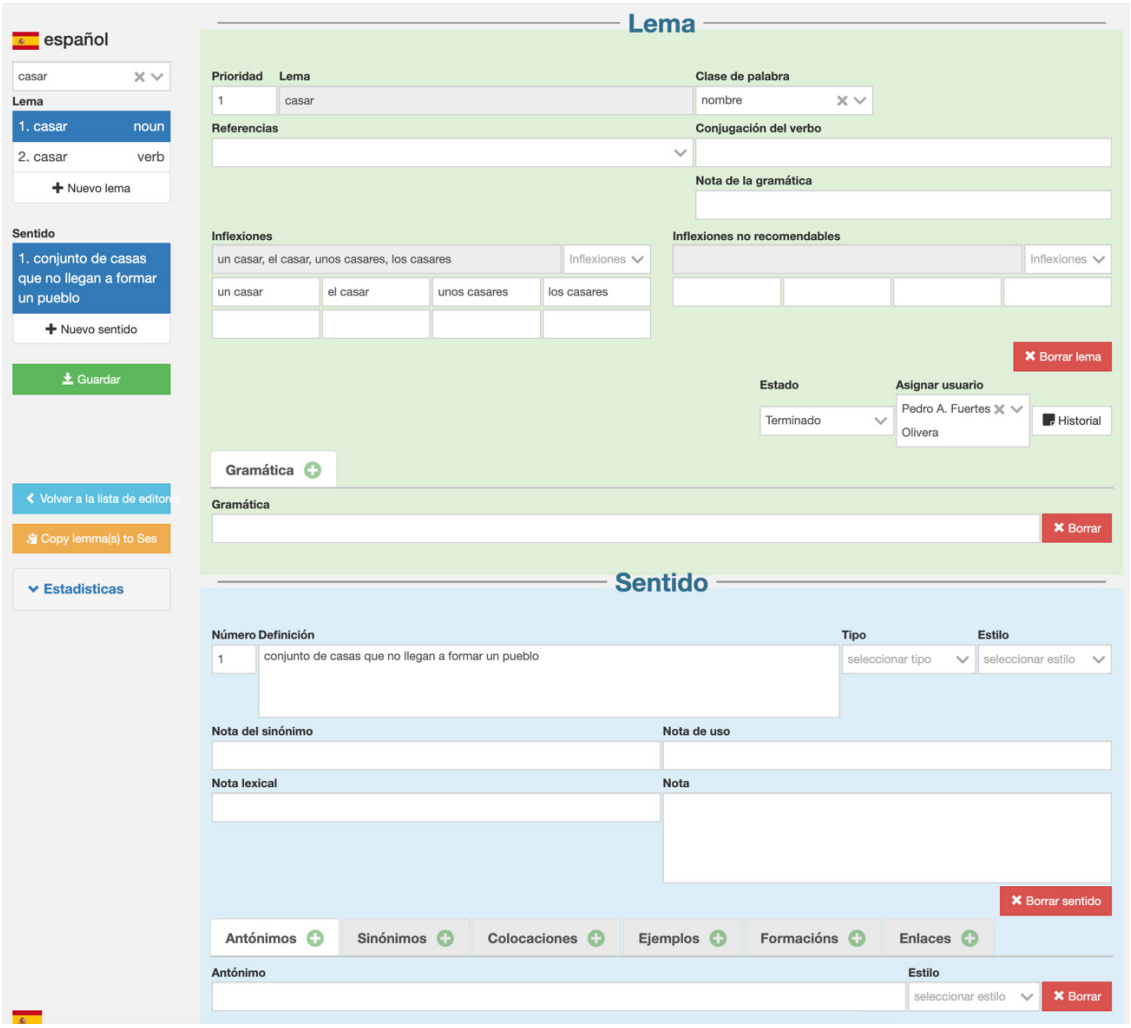
7. Innovations connected with technology

Technology is both the application of knowledge to reach practical goals and the product or service of such endeavor (Wikipedia). Online dictionaries are very different from printed ones, mostly due to the options that technology offers to lexicographers. In *DIDES,* technology has allowed us to introduce some innovations in Spanish lexicography. These are concerned with the Dictionary Writing System (DWS), the search system, and the dictionary homepage.

The DWS (it is also known as editor or lexicographic database) is the software used for writing and producing reference works such as dictionaries, glossaries, vocabularies, etc. Kilgarriff (2006: 7) states that it basically consists of an editor, a database, a Web interface, various management tools, and a kind of dictionary grammar which specifies the structure of the dictionary. Figure 3 shows a screenshot of Spanet, the DWS of *DIDES*:

Figure 3. A screenshot of the DWS for *DIDES*





Spanet is an in-house DWS, i.e., it was designed and created by the editor of the project and IT staff at Ordbogen headquarters, and consequently it suits the necessities of this project. It consists of four functionalities ("buscar notas", "entradas asignadas", "enlaces rotos", "estadísticas") and an editor or work bench:

1. "Buscar notas" is used to search for elements that were written in the slot "Nota", located in the blue part of the editor, under "Sentido". It is typically used for searching for keywords that were written by the lexicographers for highlighting any aspect connected with the compilation process or the characteristics of the lexicographic data. For instance, searching for "lemma" retrieves all the lemmas that have "low frequency in Spanish reference corpora" (section 3). Therefore, this functionality is for internal usage, is suitable for storing data, and can be used for describing the characteristics of the latter.
2. "50 entradas asignadas (en orden cronológico)" offers a list of lemmas that must be completed by the lexicographer. It also indicates when they were assigned and their word class.
3. "Enlaces rotos" informs the editor of *DIDES* of links that are not working and must be restored. The information offered is updated every night. This is a crucial innovation in *DIDES* because it has more than 20,000 links (e.g., a figure of a tree or an animal) and these need to be *always* operative.
4. "Estadísticas mensuales" simply informs how many new meanings a lexicographer has completed in a month.
5. The editor has two parts, one green and one blue. The green one contains slots for several administrative and managerial tasks and for storing the following lexicographic data: lemma, word class, inflections (recommended and not recommended), verb conjugations, grammar and grammar note. The blue one also contains slots for administrative and managerial tasks as well as for storing meanings, geographical and diastratic varieties, synonyms, antonyms, formations, links, examples and "frases", i.e., chunks of words that offer some useful information on the usage and/or function of the lemma.

The main innovations of this DWS are: (a) the technology for giving information on dead links and (b) the software used which allows the transfer of each item of data in many ways and for many different purposes; this would typically be for reusing them to create another dictionary (see below) and for selling them (section 9). For example, the data can be used for creating an initial list of 3,000 lemmas dealing with health (medicine, veterinary medicine).

The search system refers to the technological know-how used for retrieving the data stored in the DWS and shown on the dictionary homepage. *DIDES* uses four types of searches:

– == lemma, i.e., the system retrieves *exactly* the search string;
– =lemma, i.e., the system retrieves part of the lemma;
– *lemma*, i.e., the system retrieves parts of the search string;
– and fuzzy search, i.e., the system retrieves a list of results based on likely relevance.

The above types of search systems are used in three main types of innovative searches. The first type makes use of both *maximizing* and *minimizing searches*. Bergenholtz (2011: 44) indicates that a maximizing search terminates when the field or slot of the database has been explored in full, whereas in a minimizing search the system stops searching as soon as it finds what it is searching for in any field or slot of the database, each of which has been previously ordered according to lexicographic criteria.

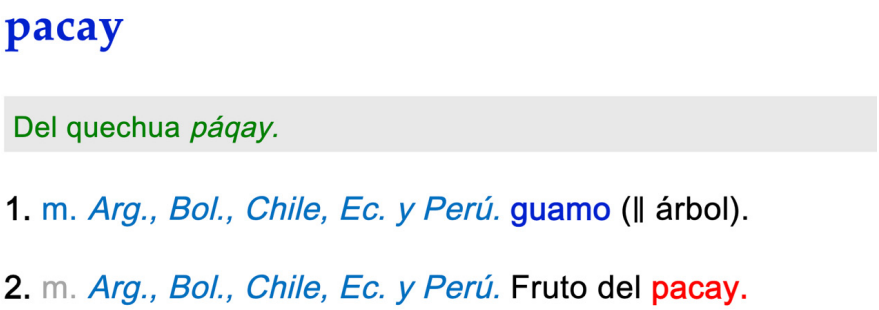
Using either one or the other search type results in different findings, two of which in *DIDES* are: *WordFinder* and *General Synonyms.* The first will allow users to retrieve by ordering something within a list of categories (section 8). For example, it can order all the figurative meanings of *DIDES* (more than 30,000), or all the extended-units-of-meaning lemmas, i.e., lemmatized expressions (more than 13,000 at the moment but expected in the future to exceed 30,000). *General Synonyms* is a tab at the top right end of the dictionary homepage that inform users of all possible meanings of a lemma in a single consultation.

The second type is necessary for retrieving multiword expressions and forms of Spanish conjugations, such as “hubieran comido”. This innovation is partly used in Spanish dictionaries such as DLE, which also allows retrieving expressions such as *a sangre fría*, but not conjugated forms such as *hubiera comido* and *habría estudiado.*

The third type allows searches in all the slots of the lexicographic database. The search can be simple, i.e., in one single slot, or multiple, i.e., combining several slots. The simple search will be used for creating lists of lexicographic data that will be accessible on a subscription basis or on demand (section 8). The multiple search can be used for creating specific types of dictionaries. For example, we can create a semi-explicative synonym dictionary, a product which, to the best of my knowledge, is not found elsewhere in Spanish lexicography. This is a production dictionary that retrieves definitions as well as synonyms and antonyms when the lemma shows homonymy or polysemy, but only synonyms and antonyms in all other situations. This type of dictionary uses a minimizing search system that connects the lemma with several slots or fields: the part of speech field; the meaning field; the synonym field; the style field of the synonym; the antonym field; the style field of the antonym; and the synonym remark field. If the lemma shows homonymy and/or polysemy, all these fields are activated, but if the lemma is monosemous, only the synonym and antonym fields are activated.

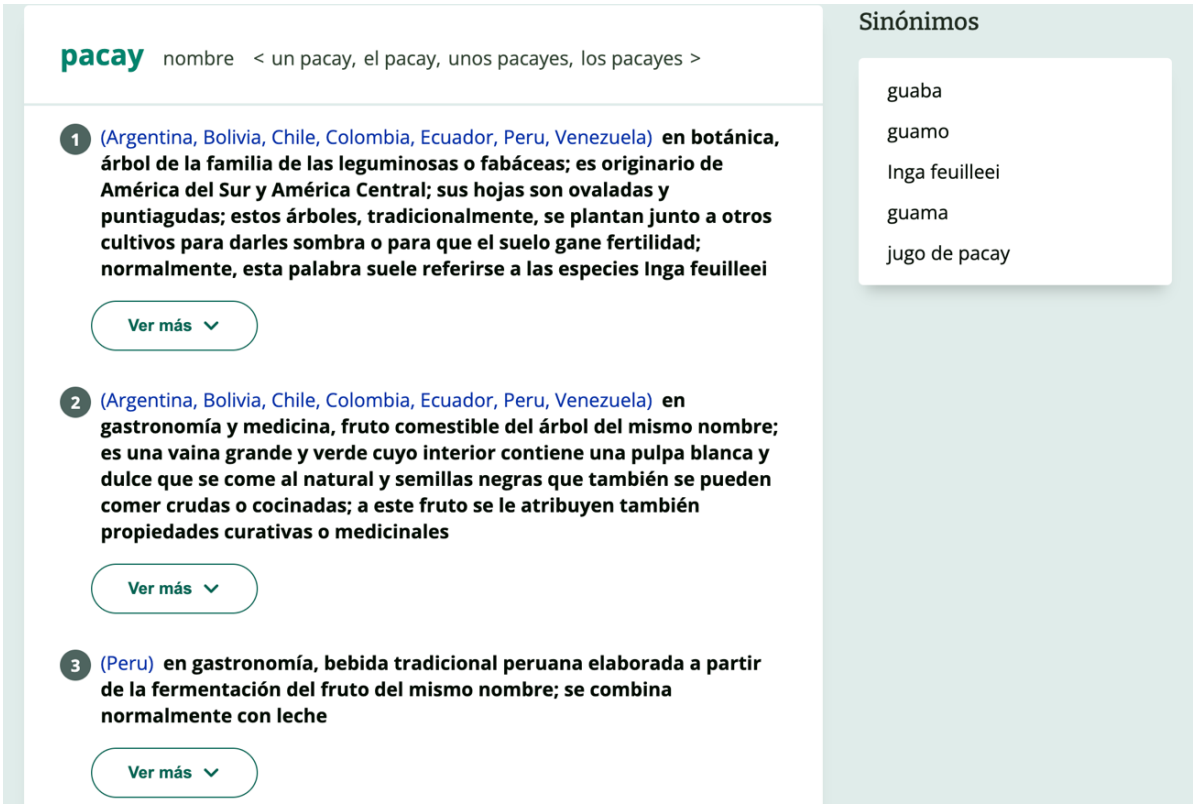
The dictionary homepage is based on the concept of simplicity and technological options. A comparison of the homepage for **pacay** in DLE (Figure 4) and *DIDES* (Figures 5 to 7) illustrates the above philosophy, which is based on the use of very advanced technology. From Figure 4, users know the following: that **pacay** derives from quechua, is a tree, is “m”, is used in some American countries, its synonym is *guamo*, and it has a fruit that is also known by the same word as the tree. Yet most users of *DLE* would not be able to answer the question “what is a pacay”? if such a question were asked.

Figure 4. The dictionary entry **pacay** in *DLE*



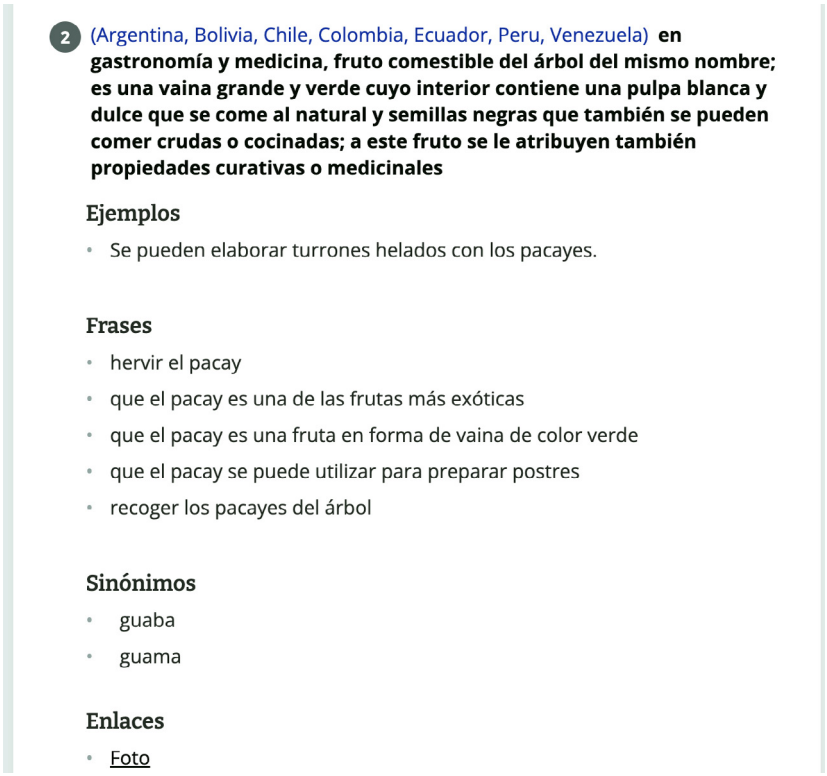
DIDES offers two options. The default option tells users what type of tree and fruit a **pacay** is, and that it is a traditional drink in Perú (their definitions), that it is a noun that goes with the articles “un, el, unos, los” and that it is used in some American countries. (Figure 5). It also shows a list of all the synonyms of **pacay**, thus offering a complete semantic picture of the lemma, an innovation that will be much appreciated by educated Spanish users who can “imagine” a complete semantic picture of the word at a glance.

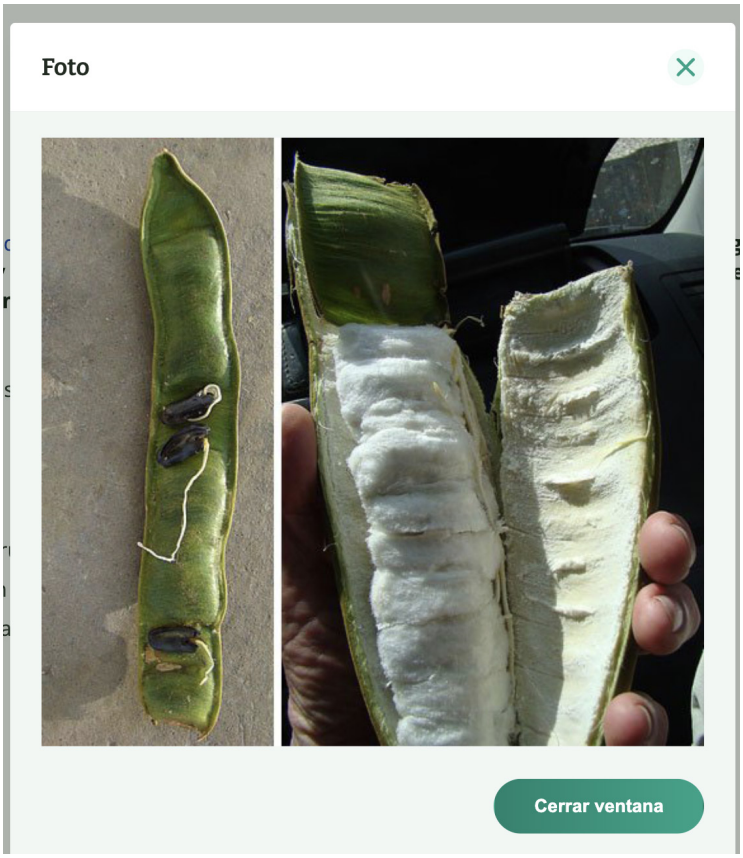
Figure 5. Dictionary articles for “pacay” in DIDES (default option)



The other option is an extended one. It is activated when the user clicks on the tab “Ver más”, which adds to the default option by displaying synonyms (and/or antonyms, if they exist), each with notes (e.g., *Inga* is the formal synonym), examples and links to external sources, e.g. a photo (Figure 6).

Figure 6. Dictionary article for “pacay” in DIDES (extended option)





These two options are especially useful for display on small screens such as those typical of smartphones. It will be possible to have all the meanings of around 85% of the lemmas without scrolling down the screen.

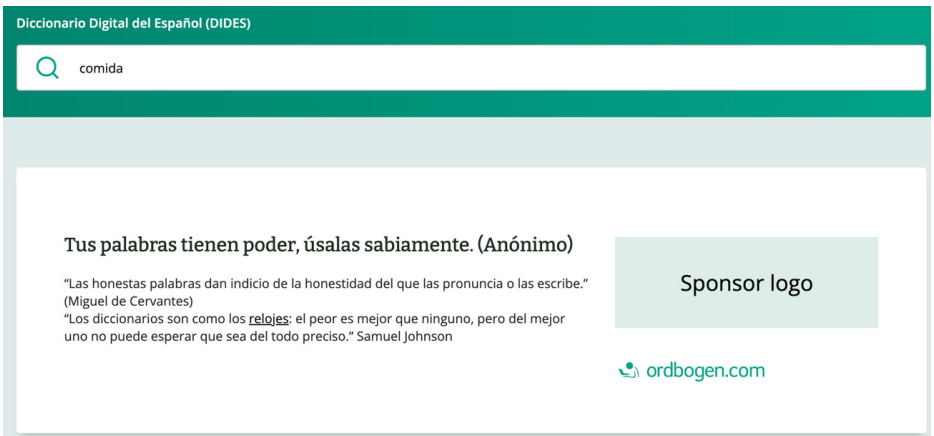
8. Innovations connected with Business Models

Since 2014, Spanish speakers have had free access to several dictionaries designed and maintained by the "Reales Academias" of Spanish speaking countries, especially DLE. This has had profound consequences in the "lexicography industry", i.e., the research and business activities connected with theoretical and practical lexicography. Since then, it can be said that the private lexicographic sector is practically non-existent (e.g., publishing houses have closed their lexicographic units and no new dictionary of general Spanish has been published in any format); also, that the public sector (i.e., that depending on public funds for research into lexicography and its formalization in the shape of real dictionaries) is somewhat chaotic, with no one knowing which project is or is not financed and with apparently no long-term view envisaged (it seems that the Research Agency prefers financing "prototypes" instead of more consolidated projects).

Sustainable lexicography, then, needs fresh ideas. Thus, *DIDES* has been prepared for:

1. Facilitating sponsorships. Figure 7 shows the homepage of *DIDES,* which reserves rooms for one or more sponsors:

Figure 7. Homepage of *DIDES*



2. Facilitating its location and retrieval in the digital world. For instance, googling "DIDES" + Diccionario, or "Diccionario Digital del Español" retrieves the dictionary in the first or second position of search.
3. Preparing a specific publicity campaign for it, highlighting its advantages against existing dictionaries. For instance, FIAPE (Federación Internacional of Asociaciones de Profeesores de Español) has posted information on DIDES in its homepage (see: https://fiape.org/?p=3766)
4. Preparing the lexicographic data stored in the DWS for multiple potential usages, e.g. for selling the definitions for making bilingual dictionaries. For instance, example (8) shows a DIDES definition and its translation into English with DeepL Translator, a neural machine translation service whose "algorithm uses convolutional neural networks and an English pivot" (DeepL Translator: Wikipedia).

habladuría
palabrería, es decir, empleo de muchas palabras que no dicen nada pero que suenan muy bien; se hace para presumir o impresionar
chatter
chatter, i.e. the use of many words that say nothing but sound very well; it is done to show off or to impress

Example (8): Definitions in *DIDES* and its translation with DeepL Translate.

This example and the launch of generative AI such as ChatGPT indicates that Artificial Intelligence can play key roles in lexicography, an idea that merits further investigation and which will address in upcoming papers.

9. Conclusion

This article has discussed some of the main innovations of *DIDES,* an online general dictionary of Spanish that is part of the lexicographic project "Diccionarios Valladolid-UVa". These innovations concern all aspects of dictionary making, from selecting lemmas to finding out data regarding their meanings and usages. The following features are especially relevant:

- Lemmas are selected in three related steps: (a) an initial lemma list extracted from log-files; (b) amplification of the initial lemma list based on grammar rules and systematized lists; (c) continuous updating.
- Internet is the main lexicographic source, although existing dictionaries, grammar books, and other reference works, e.g., usages, are also been consulted.
- It pays attention to the linguistic and social environment of all Spanish-speaking countries and aims at offering linguistic data from this broad sector.
- It does not force users to refer to several dictionary articles, e.g., it does not use recursive definitions. Everything is simple and aims to eliminate data and information overload (Gouws and Tarp, 2017).
- It generally uses new and in-house technology, e.g., by means of a very flexible editor which allows lexicographers and IT staff to offer users different options. Furthermore, the technology used allows the on-going process of updating dictionary articles, i.e. their continuous updating without waiting for new editions. To sum up, editions are no longer necessary as the dictionary is continually changed, modified and updated. For instance, any modification of the data stored in the DWS is visible one second after the editor of the project saves the changes made.
- It proposes a business model based on offering high-quality data which can be easily published and sold on demand.

Acknowledgment

Special thanks are due to Prof. Sven Tarp, Aarhus University, for his constructive comments.

References

Bergenholtz, Henning & Sven Tarp. (2003). Two opposing theories: On H.E. Wiegand's recent discovery of lexicographic functions. *Hermes, Journal of Linguistics* 31, 171-196.

Colman, Lut. (2016). Sustainable lexicography: Where to go from here with the *ANW* (*Algemeen Nederlands Woordenboek*, an Online General language Dictionatry of Centemporary Dutch? *International Journal of Lexicography* 29 (2): 139-155. https://doi.org/10.1093/ijl/ecw008

CREA: Corpus de Referencia del Español Actual. Real Academia Española de la Lengua. https://corpus.rae.es/creanet.html (Last consultation: January, 2024)

CORPES XXI: Corpus del Español del Siglo XXI. Real Academia Española de la Lengua. https://www.rae.es/banco-de-datos/corpes-xxi (Last consultation: January, 2024).

DeepL Translator: https://en.wikipedia.org/wiki/DeepL_Translator (Last consultation: January, 2024).

DEL: Diccionario de la Lengua Española. RAE. https://dle.rae.es/ (Last consultation: January, 2024)

DIDES. Diccionario Digital del Español. Centro Internacional de Lexicografía de la UVa. Odense: Ordbogen. https://diesgital.com/ (Last consultation: January, 2024).

Forster, Kenneth I & Chambers, Susan M. (1973). Lexical access and naming time." *Journal of Verbal Learning and Verbal Behavior* 12(6): 627-635. https://doi.org/10.1016/S0022-5371(73)80042-8.

Fuertes Olivera, Pedro A. (1992). *Mujer, lenguaje y sociedad. Los estereotipos de género en inglés y en español*. Madrid: Ayto. de Alcalá de Henares.
Fuertes-Olivera, Pedro A. (2019). Designing and making commercially driven integrated dictionary portals: The *Diccionarios Valladolid-UVa.*" *Lexicography* 5: 1-21. https://doi.org/10.1007/s40607-019-00056-8
Fuertes-Olivera, Pedro A. (2022). The mental lexicon in lexicography: The *Diccionarios Valladolid-UVa*. *Lexikos* 32 (1): 118-140. https://doi.org/10.5788/32-1-1712
Fuertes-Olivera, Pedro A., Sven Tarp & Sepstrup, Peter. (2018). New insights in the design and compilation of digital bilingual lexicographical products: the case of the *Diccionarios Valladolid-UVa*. *Lexikos* 28: 152-176. https://doi.org/10.5788/28-1-1460
Fuertes-Olivera, Pedro A. & Tarp, Sven. (2020). A window to the future: Proposal for a lexicographically-assisted writing assistant". *Lexicographica* 36: 257-286. https://doi.org/10.1515/lex-2020-0014
Fuertes-Olivera, Pedro A. & Tarp, Sven. (2022). Critical lexicography at work: Reflections and proposals for eliminating the gender bias in general dictionaries of Spanish. *Lexikos* 32(2): 105-132: 105-132. https://doi.org/10.5788/32-2-1699
Fuertes-Olivera, Pedro A. A (2025). *Guide to Practical Online Lexicography*. London: Routledge.
Gouws, Rufus & Tarp, Sven. (2017). Information overload and data overload in lexicography. *International Journal of Lexicography* 30 (4): 389-415. https://doi.org/10.1093/ijl/ecw030
Holmes, Janet & Meyerhoff, Miriam (Eds.). (2003). *The Handbook of Language and Gender.* Oxford: Blackwell. Innovation: https://en.wikipedia.org/wiki/Innovation (last consultation: January 2024).
Kilgarriff, Adam. (2006). Word from the chair. In G_M de Schryver (ed.) *DWS. Proceedings of the Fourth International Workshop on Dictionary Writing Systems*. Pretoria: SF Press, 7.
Nissen, Uwe K. (1986). Sex and gender specifications in Spanish. *Journal of Pragmatics* 10: 725-738.
Tarp, Sven. (2008). *Lexicography in the Borderland Between Knowledge and Non-knowledge*. Tübingen: Niemeyer.
Tarp, Sven. (2011). Lexicographic and other e-tools for consultation purposes: Towards the individualization of needs satisfaction, in Pedro A. Fuertes-Olivera & H. Bergenholtz (eds.), *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. London: Continuum, 54-70.
Tarp, Sven. (2022). Turning bilingual lexicography upside down: Improving quality and productivity with new methods and technology*." Lexikos*, 32: 66-87. https://doi.org/10.5788/32-1-1686
Tarp, Sven & Fuertes-Olivera, Pedro A. (2016): Advantages and disadvantages in the use of internet as a corpus: The case of the online dictionaries of Spanish Valladolid-UVa. *Lexikos* 26: 273-295. https://doi.org/10.5788/26-1-1349