

Making Lexicography Sustainable: Using ChatGPT and Reusing Data for Lexicographic Purposes*

Pedro A. Fuertes-Olivera, *Department of Afrikaans and Dutch, University of Stellenbosch, South Africa; International Centre for Lexicography, University of Valladolid, Spain; and Centre of Excellence in Language Technology, Ordbogen A/S, Odense, Denmark*
(*pedro@emp.uva.es*) (<https://orcid.org/0000-0003-3831-5377>)

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, 2022a, 2022b; Fuertes-Olivera et al. 2018; Tarp and Fuertes-Olivera 2016). Each partner gave around €180,000 (the International Centre for Lexicography's contribution came from several research projects funded by the Spanish Research Agency), 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). The above project has produced several results, with the recent publication of the *Diccionario Digital del Español* (DIDES) its most relevant result (<https://diesdigital.com>). Within the framework of these projects, this paper offers a general introduction of the project (Section 1), refers to the concept of sustainable lexicography (Section 2), indicates that sustainability lexicography implies a better understanding of lexicographic data (Section 3), and increasing lexicographic productivity, e.g., by crafting definitions for AI translations (Section 4) and using generative AI chatbots such as ChatGPT in the day-to-day lexicographic work.

Keywords: CHATGPT, DEEPL TRANSLATE, DICCIONARIOS VALLADOLID-UVA, LEXICOGRAPHIC PRODUCTIVITY, SUSTAINABLE LEXICOGRAPHY, PUBLIC FUNDING, GENERATIVE AI

Opsomming: Hoe om die leksikografie volhoubaar te maak: Die gebruik van ChatGPT en die hergebruik van data vir leksikografiese doeleinades. In 2014 het die Internasionale Sentrum vir Leksikografie, 'n navorsingsgroep by Valladolid, 'n kontrak vir die ontwikkeling van 'n leksikografiese projek, die sogenaamde *Diccionarios Valladolid-UVA*, met Ordbogen A/S ('n Deense taaltegnologiemaatskappy) en die Universiteit van Valladolid onder-

* A version of this paper was presented at StellenLex 2024: Lexicography, Artificial Intelligence, Language Models and Innovation, held at the Bureau of the Woordeboek van die Afrikaanse Taal on January 24, 2024.

teken (Fuertes-Olivera 2019, 2022a, 2022b; Fuertes-Olivera et al. 2018; Tarp en Fuertes-Olivera 2016). Elke vennoot het ongeveer €180,000 bygedra (die Internasionale Sentrum vir Leksikografie se bydrae was afkomstig van verskeie navorsingsprojekte wat deur die Spaanse Navorsingsagentskap befonds is) wat gebruik moes word in die ontwerp en samestelling van Spaanse woordeboeke (spesifieker 'n algemene Spaanse woordeboek, 'n Spaanse rekeningkundewoordeboek, 'n tweetalige Spaans-Engels/ Engels-Spaanse woordeboek en 'n tweetalige Spaans-Engels/Engels-Spaanse rekeningkundewoordeboek. Bogenoemde projek het verskeie resultate tot gevolg gehad, met die onlangse publikasie van die *Diccionario Digital del Español* (DIDES) as die mees relevante produk (<https://diesgital.com>). Binne die raamwerk van hierdie projekte verskaf dié artikel 'n algemene inleiding tot die projek (Afdeling 1), word daar verwys na die konsep van volhoubare leksikografie (Afdeling 2), en word daar aangege toon dat volhoubare leksikografie 'n beter begrip van leksikografiese data (Afdeling 3), toenemende leksikografiese produktiwiteit, bv., deur die skep van definisies vir KI-vertalings (Afdeling 4), en die gebruik van generatiewe KI-kletsbotte soos ChatGPT in daaglikske leksikografiese take impliseer.

Sleutelwoorde: CHATGPT, DEEPL TRANSLATE, DICCIONARIOS VALLADOLID-UVa, LEKSIKOGRAFIESE PRODUKTIWITEIT, VOLHOUBARE LEKSIKOGRAFIE, OPENBARE BEFONDSING, GENERATIEWE KI

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 language technology 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 four part-time lexicographers, each with a 19-hour 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 four of the ten applicants. These four selected lexicographers started their work in March 2014; they all worked for four hours from Monday to Thursday and three hours on Friday. They were in the same room, next to the office of the editor of the project, who could check their progress and answer their queries very easily and quickly. They worked on the project until June 2020, when the Spanish Research Agency decided to stop funding the research projects they had been financing up to that time.

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. The editor is totally committed to

creating more dictionary articles for the general dictionary of Spanish, while also being open and committed to adapting the existing dictionary articles to incorporate new ideas and technological possibilities, and together with Sven Tarp, to explain the decisions taken and to explore new theoretical and practical possibilities in lexicography. It is assumed that these are truly innovative possibilities, i.e., they are the result of the development of more effective products, services, processes, technologies, and business models. Tarp (2022), for example, refers to a current project he is involved in that substantially modifies the concept of bilingual lexicography and opens the room for the use of generative AI chatbots in several lexicographic activities (see Section 4).

This article assumes that sustainable lexicography cannot be achieved without proper and regular funding and a true and effective analysis of the results obtained with the funds received. This implies a better understanding of the concept of sustainable lexicography.

2. The concept of sustainable lexicography

Sustainability in lexicography generally refers to the working conditions, re-using of lexicographic material, and financial resources that are needed for designing, making, and maintaining any lexicographic project. Kosem et al. (2021), for example, defend that semantic data should no longer exist in isolation and propose different ways for managing large, interconnected datasets. They assume that the different projects on data consolidation currently in operation will have an impact on both theoretical and practical lexicography, e.g., the outcome of the European Lexicographic Infrastructure (ELEXIS) project, which is a collaborative initiative aimed at fostering innovation and cooperation in the field of lexicography across Europe. Tiberius et al. (2024), for instance, describe the results of three international surveys that were carried out in the context of ELEXIS and that aimed at gaining insight into lexicographic practices and the lexicographers' needs in Europe.

One of the main objectives of the ELEXIS project is to integrate and make accessible the rich lexicographic resources of Europe, including dictionaries, lexical databases, and related linguistic tools and datasets. Without any doubt, such integration will reduce lexicographic costs in time and funds, and thus will make sustainable lexicography possible, especially by shifting "towards open access structured data enabling re-use and linking of dictionary data along with stand-alone lexicographic (and terminological) resources into numerous dictionary portals." (Tiberius et al. 2024: 23)

At a more down-to-earth level, lexicographic projects have to face specific drawbacks, e.g., lack of funds. Colman (2016) describes a lexicographic project (*The Algemeen Nederlands Woordenboek* (ANW), *Dictionary of Contemporary Dutch*); this project was in a similar situation to the project *Diccionarios Valladolid-UVa*; several partners initially allocated funds for the projects, but when the partners

decided to stop funding them, they themselves had to find their own resources for continuing.

Colman (2016: 140-141) describes the ANW project as an online dictionary "through which a range of users can explore the Dutch vocabulary" and as a "linguistic data resource from which users, and especially language professionals, can extract data necessary for their research." Colman's (2016) distinction between dictionaries and lexicographic data is based on an economic and environmental interpretation of *sustainability* which demands, *inter alia*, "reuse of materials and products", "economic use of resources", "workflow optimization" and "the weighing of costs and benefits to present and future generations." The translation of the above ideas into lexicography implies that lexicographers "will need to convince funders that their investments are not a waste of time and money and that it is possible to optimize the workflow through responsible use of materials, products and financial resources" (Colman 2016: 141). In practical terms, her concept of *sustainable lexicography* implies the following:

- reusing the content of existing dictionaries, for example, adapting existing definitions to new situations;
- using links to external data, for example to a Wikipedia page;
- reusing the data of existing Dictionary Writing Systems, for example, from a monolingual dictionary to a bilingual one;
- increasing the automation of the lexicographic process itself, for example, finding "good examples" in a corpus;
- storing as much data as possible in the lexicographic database, but adapting the presentation of the data to the usage situation and user's needs (i.e., the creation of dynamic dictionary articles (Fuertes-Olivera and Bergenholz 2011; Tarp 2011);
- making the lexicographic database usable for different purposes;
- innovating as much as possible, as shown below.

Colman (2016: 142-151) mentions four innovations in the ANW. Firstly, the traditional lexicographic definitions are complemented by a "semagram", which is basically a system of 'slots' and 'fillers' that includes all the defining characteristics of the lemma. Colman (2016: 143) claims that semagrams such as that of Table 1 (she adapts it from Moerdijk et al. 2008: 19) are useful because they enable lexicographers to make much better definitions whose additional information can "help to optimize onomasiological searches" in online dictionaries.

Secondly, the ANW offers lexicographic treatment of "combinatorics" and "phraseology" (Colman 2016: 144). These basically include "free combinations", "semi-fixed collocations", "fixed expressions" and "proverbs". They and the information for their lexicographic treatment is taken basically from corpora and retrieved by means of word sketches and collocation lists from the Sketch Engine; it aims at offering users "structured collocational information", i.e., "the combinations in real language use, mostly of binary combinations such as (a) noun + verb, (b) verb, verb + noun, (c) adjective + noun, and (d) adjective + *to* + verb":

this treatment will allow users, say, to find out "which verbs take *kat* (cat) as their subject and which verbs take *kat* as their object" (ibid. 145).

UPPER CATEGORY:	is an animal
CATEGORY:	is a bovine (animal)
SOUND:	moos/lows, makes a sound that we imitate with a low, long-drawn 'boe'
COLOUR:	is often black and white spotted, but also brown and white spotted, black, brown or white
SIZE:	is big
PARTS:	has an udder, horns and four stomachs: paunch, reticulum, third stomach, proper stomach
BUILD:	is big-boned, bony, large-limbed in build
FUNCTION:	produces milk and (being slaughtered) meat
PLACE:	is kept on a farm; is in the field and in winter in the byre
AGE:	is adult, has calved
PROPERTY:	is useful and tame; is considered as a friendly, lazy, slow, dumb, curious, social animal
SEX:	is female
BEHAVIOUR:	grazes and ruminates
TREATMENT:	is milked every day; is slaughtered
PRODUCT:	produces milk and meat
VALUE:	is useful

Table 1: Semagram for *cow*. Source: Colman (2016: 142)

Thirdly, the database of the ANW "functions as a kind of wordnet. For each word or word group in a particular sense, it includes related words such as hyperonyms, synonyms, antonyms, andronyms and feminines" (ibid. 146). She adds that some pragmatic information may be added, if necessary, as some research (e.g. Murphy 2013) has found that some users want more information about possible differences among synonyms, especially differences in connotation and linguistic variety. She also acknowledges that *wordnets* are difficult to process, structure and present in a dictionary. The ANW has used the thesaurus function of Sketch Engine for registering lexical and grammatical relations and includes meaning relationships "like metaphor, metonymy, generalization and specialization" when relevant (ibid. 149).

Finally, the ANW includes a large list of "simplexes", i.e., derivatives and compounds (ibid. 149), some of which are difficult to spell and some of which demonstrate the existence of regularities in word formation.

Colman (2016) mentions several drawbacks or weaknesses in each of the innovations she discusses. My view of these is mixed, as I also use some of the above ideas in the *Diccionarios Valladolid-UVa* (for example, the lemmatization of multi-word lemmas; see Fuertes-Olivera 2019 and 2022a), but I also find drawbacks that are not mentioned or assumed as such. Firstly, all the innovations discussed are language-centered, i.e., they assume that dictionaries are language artefacts and that "the art and craft of dictionary making" can be solved by offering users better and more language data. My view is that lexicographic data is much more than language data and need a proper understanding of its nature and possible functions (see Section 3). Secondly, the innovations proposed must be also analyzed in terms of lexicographic productivity, especially in terms of the money and time spent for creating the lexicographic data. For example, the application of the concept of "semagram" will be very time consuming and mostly useless as it cannot be easily implemented with many lemmas, especially with verbs, adjectives, adverbs, conjunctions, intangible nouns and so on. Instead, semagrams such as that of "cow" can be substituted by a figure and/or by using definitions from chatbots, i.e., the use of existing technology for speeding up the lexicographic process and reducing costs (see Section 4).

3. The concept of lexicographical data

Lexicographic data are typically defined as any data that have been prepared or accepted by lexicographers and stored in a Dictionary Writing System (DWS) with the aim of helping humans and/or machines convert them into information in a *straightforward manner* (Fuertes-Olivera et al. 2018; Fuertes-Olivera and Tarp 2020). Lexicographic data can be economic resources (and hence, contribute to the concept of sustainable lexicographic) assuming that:

- They are presented in any format, e.g., as words, figures, sounds, drawings, symbols, running texts, etc.
- They may have been prepared by the lexicographers themselves or by someone else (the possibility of linking external data); this increases the "offer" of data, reduces lexicographic costs, and emphasizes that lexicographers must work with more than linguistic data.
- They must be crafted for converting them into information in a single cognitive process. This is a crucial point in our definition of lexicographic data. In these circumstances, most data in, say, existing Spanish dictionaries are not lexicographic, as they cannot be understood due to several flaws in their treatment and presentation, especially in terms of the use of a compact and traditional lexicographic style full of abbreviations, recursive definitions, and scarce relevant data (Nomdedeu-Rull and Tarp 2024).

Figure 1 shows the dictionary article **pacay** in the *Diccionario de la Lengua Española* (DLE):

pacay

Del quechua *páqay*.

1. m. *Arg., Bol., Chile, Ec. y Perú. guamo* (|| árbol).
2. m. *Arg., Bol., Chile, Ec. y Perú. Fruto del pacay.*

Figure 1: The dictionary article **pacay** in the DLE

Figure 1 only informs that:

- it is a tree, whose fruit is also called "pacay";
- the tree is also called "guamo";
- it is used in two countries (Chile and Perú), and in three others identified as "Arg.", "Bol." and "Ec.>";
- it derives from "quechua";
- it is "m."

In other words, only human users who already knew what a "pacay" is can convert the data of the article into information. Such an article shows that the DLE is a "faster horse" (Tarp 2011), i.e., a printed dictionary with digital access that has not been adapted to the digital medium (Bergenholtz et al. 2009; Fuertes-Olivera 2018; Fuertes-Olivera and Bergenholtz 2011; Fuertes-Olivera and Tarp 2014; Granger and Paquot 2012). In sum, the creation of such data will make lexicography unsustainable and should rather be avoided.

DIDES uses a different approach as seen in Figure 2.

Figure 2 offers the following data about **pacay**:

- a noun with three senses — a tree, the fruit of the tree, and a traditional Peruvian drink. The tree belongs to the *Fabaceae* or *Leguminosae* family, comes from South and Central America; its leaves are oval; it is usually planted for shading other crops and fertilizing soils. The fruit is a green and big edible sheath with black seeds that can also be eaten or used in traditional medicine. The drink is typically combined with milk;
- **pacay**, and **pacayes** are its singular and plural forms;
- its accompanying articles are "un", "el", "unos", and "los";
- it has synonyms, each with its diastratic information;
- it offers examples of the three meanings used in several contexts;

- it is used in five countries: Argentina, Bolivia, Chile, Ecuador and Perú;
- it offers links to figures, e.g., to the fruit;
- it includes the buttons "ver más" (see more) or "ver menos" (see less) for accommodating the data to the size of the screen;
- it offers a difference between the tree (it is used in botany), the fruit (it is used in gastronomy and medicine), and the traditional Peruvian drink.
- It offers a complete set of clickable synonyms, which offer a complete semantic picture of the lemma, and favor cross-referencing.

pacay nombre < un pacay, el pacay, unos pacayes, los pacayes >

1 (Argentina, Bolivia, Chile, Colombia, Ecuador, Peru, Venezuela) en botánica, árbol de la familia de las leguminosas o fabáceas; es originario de América del Sur y América Central; sus hojas son ovaladas y puntiagudas; estos árboles, tradicionalmente, se plantan junto a otros cultivos para darles sombra o para que el suelo gane fertilidad; normalmente, esta palabra suele referirse a las especies *Inga feuilleei*

[Ver más](#)

2 (Argentina, Bolivia, Chile, Colombia, Ecuador, Peru, Venezuela) en gastronomía y medicina, fruto comestible del árbol del mismo nombre; es una vaina grande y verde cuyo interior contiene una pulpa blanca y dulce que se come al natural y semillas negras que también se pueden comer crudas o cocinadas; a este fruto se le atribuyen también propiedades curativas o medicinales

[Ver más](#)

3 (Peru) en gastronomía, bebida tradicional peruana elaborada a partir de la fermentación del fruto del mismo nombre; se combina normalmente con leche

[Ver más](#)

Sinónimos

guaba
guamo
Inga feuilleei
guama
jugo de pacay



Figure 2: The lemma **pacay** in DIDES

In sum, Figure 2 illustrates the concept of lexicographic data and its economic potential. Firstly, it can be used by humans and machines; Secondly, it really informs on meanings, forms, and functions. Thirdly, all the data can be prepared for different usages, extracted individually, and sold/licensed to third parties. Fourthly, it illustrates that lexicographic data is different from linguistic data. Finally, the use of figures, video clips and audio files, etc. may save lexicographic time and highlights the relevance of lexicographic productivity.

4. The concept of lexicographic productivity

Tarp and Fuertes-Olivera (2016) and Fuertes-Olivera (2019) have defined lexicographic productivity in economic terms; it refers to the rate at which lexicographic data are produced per unit of time, labor, or resources. For example, the lexicographic team working in the *Diccionarios Valladolid-UVa* project crafted around 35,000 dictionary articles per year for a cost of around €100,000 a year (around €3,50 per dictionary article). The number of dictionary articles crafted can be substantially increased and their cost reduced by (a) concentrating on the real nature of the lexicographic work (e.g., it is a waste of time to make a specialized dictionary without experts working on it, and (b) using existing technology for reusing existing lexicographic data and crafting new one.

The two above ideas go hand in hand and are illustrated below. For instance, the semagram in Table 1 (it can cost around one hour of work) can be substituted by asking ChatGPT for a definition of *cow* (below in example 1; it takes half a minute or less):

A cow, scientifically known as Bos taurus, is a domesticated mammal belonging to the family Bovidae. Cows are one of the most common types of livestock, raised globally for their ability to produce milk, meat (beef and veal), and hides, which are used in leather manufacturing. The term "cow" specifically refers to female members of the species that have given birth to at least one calf, whereas males are known as bulls, and young cattle are called calves. Cows are herbivores, primarily grazing on grasses, and they have a unique digestive system with a four-chambered stomach, which allows them to break down tough plant materials through a process called rumination.

Cows play a significant role in agriculture and have been domesticated for thousands of years, serving not only as a source of various commodities but also as draft animals in many cultures. Beyond their economic value, cows hold profound cultural and religious significance in various societies. In some cultures, cows are revered as sacred animals and symbolize fertility, motherhood, and abundance. Their importance in human history, agriculture, and culture makes them indispensable to societies worldwide.

Example 1: Definition of **cow** generated with ChatGPT (February 16, 2024).

Lexicographically speaking, example 1 offers a lot of information for a human lexicographer to create dictionary articles. At least, five lexicographic defin-

tions can be crafted from the ChatGPT definition in a Spanish dictionary (the first three are in the DLE and all five in the DIDES:

1. A cow is a female herbivore mammal; it is typically found in farms and ranches throughout the world; it is raised for meat, milk, or leather, is docile and is easy to work with.
2. Cow is also a type of meat that is consumed by humans.
3. Cow is also a type of leather. This should be modified (recreated by the human lexicographer) by indicating that it is not the cow but its skin which is tanned and then used in the textile industry.
4. Cows are docile and big animals. Hence, they can be used metaphorically to refer to fat and dumb people. This meaning is informal and much used in Spanish (it is surprising that it is not included in DLE).
5. Cows are cultural and religious symbols in some societies.

In addition to the above data, example 1 also indicates that its male counterpart is called a "bull", and that young animals are "calves".

Example 1 shows that the introspection and knowledge of a well-trained human lexicographer working with generative AI chatbots such as ChatGPT, Google searches (Google minitexts in Tarp and Fuertes-Olivera 2016), log files, and technology for reusing lexicographic data may enhance substantially lexicographic productivity, thus reducing costs and making the production of dictionaries cheaper. In my view, this practice is much better than working with concordances, key words, and other corpus-based or -driven technologies. Three examples illustrate this idea.

Firstly, Tarp (2022: 68), for example, is working in a project based on two experiments:

1. Using artificial intelligence to select adequate example sentences and automatically assign them to the relevant senses in a lexicographical database.
2. Using machine translation to translate L2 definitions into L1, where the translated definitions can both explain the meaning of L2 lemmata and functions as semantic differentiators when bridging from L1 to L2.

Regarding the use of Artificial Intelligence, Tarp and Henrik Hoffmann (an IT expert working at Ordbogen A/S) translated 200 Spanish definitions of the Spanish monolingual dictionary of the *Diccionarios Valladolid-UVa* project into English with the help of Google Translate and DeepL Translate (two AI-based translation tools). They found out that 78% of those translated with DeepL Translate were correct and did not need any more intervention by a lexicographer. We (Tarp, Hoffmann and myself) discussed the results and observed that automatic translations improved if the Spanish definitions were crafted adding the defining features of the *definiens* without changes of flow (e.g., without inserting non-defining relative sentences for clarifying features of the lemma being defined), using simple and clear clauses (e.g. without using subjunctives and long sentences), separating the defining features by semi-colons (instead of

stops and non-defining relative clauses), and contextualizing them. Example 2 shows the legal definition of **bancarrota** (bankruptcy) in DIDES and its translation with DeepL Translate before we studied them:

- **bancarrota**

en derecho, situación legal declarada por un juez; consiste en hacer perder a una persona, empresa, institución, organismo, etc. la disposición y administración de sus bienes, restringir su capacidad e inhabilitarle para el ejercicio de la actividad económica

- **bankruptcy**

in law, a legal situation declared by a judge; it consists of making a person, company, institution, organization, etc. lose the disposition and administration of its assets, restricting its capacity and disqualifying it from exercising economic activity

Example 2: Initial translation of **bancarrota** with DeepL Translate

Example 2 shows that the defining features of **bancarrota** are separated by commas and semicolons. These are: (a) it is a legal term; (b) the situation occurs when a judge declares it; (c) proprietors of the asset lose it; (d) proprietors cannot continue administering the asset; (e) proprietors cannot continue with the same economic activity. The first four characteristics are perfectly translated; the fifth one, however, may be wrongly translated because the Spanish original uses "inhabilitarle" (the verb goes with a singular clitic referring a person). This can be easily corrected, e.g., by using the plural instead of the singular, as shown in example 3. The new translation is totally correct and has two interesting modifications: (a) "their" is used instead of "its" in "lose the disposition and administration of their assets" and (b) "them" is used instead of "it" in "disqualifying them":

- **bancarrota**

en derecho, situación legal declarada por un juez; consiste en hacer perder a una persona, empresa, institución, organismo, etc. la disposición y administración de sus bienes, restringir su capacidad e inhabilitarlos para el ejercicio de la actividad económica

- **bankruptcy**

in law, a legal situation declared by a judge; it consists of making a person, company, institution, organization, etc. lose the disposition and administration of their assets, restricting their capacity and disqualifying them from exercising economic activity

Example 3: Modified translation of **bancarrota** with DeepL Translate

Secondly, the *Diccionarios Valladolid-UVa* has also two own technologies (they were created by IT staff at Ordbogen under lexicographers' guidance) for reusing data. One of them is a "copy to" button in the DWS of the general monolingual dictionary (Figure 3; the orange button):

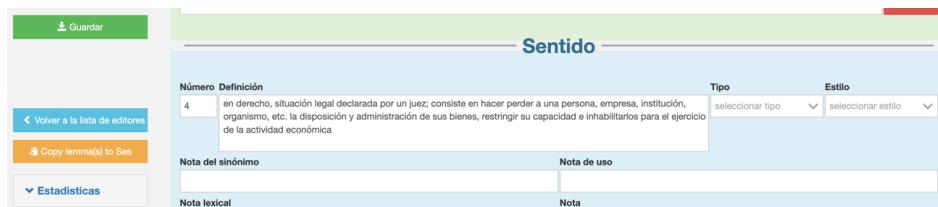


Figure 3: The "copy to" button in the DWS of the Spanish general dictionary

This button allows reusing existing lexicographic data by transferring it to another DWS. The data is transferred, for example, to the Spanish side of the bilingual Spanish–English/English–Spanish dictionary of the project. It transfers the data that are correct for all types of dictionaries, typically the definition, grammar, example(s) and links of a particular meaning. It does not reuse the lemma list as each one is based on different criteria, e.g., around 80% of the lemma list of a specialized dictionary consists of multi-word lemmas, as the lemma **just-in-time inventory control system** in accounting.

The other technology in the *Diccionarios Valladolid-UVa* is a "tool" for matching lemmas with their meanings, collocations and examples in a bilingual English–Spanish/Spanish–English accounting dictionary (Figure 4):

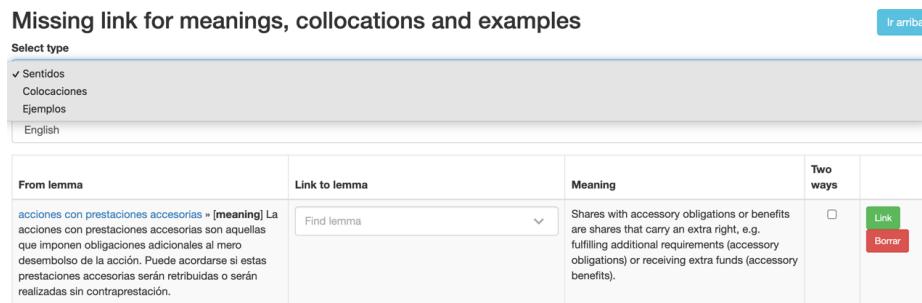


Figure 4: Tool for linking missing meanings, collocations and examples

The tool searches in a database with around 6,000 accounting lemmas, 25,000 phrases (collocations) and more than 7,000 examples in English and/or Spanish

that are "free", i.e., they are not found as dictionary articles. They were created by a team of lexicographers and accounting experts in Denmark and Spain for a printed English-Spanish dictionary of accounting, published in 2010 (Fuertes-Olivera et al. 2010).

By clicking on the blue string of words or placing them in the search button "Link to lemma" (see Figure 4), the system automatically searches for unmatched data and, when found, creates the dictionary article (or part of it) in the language searched for and stores it in its corresponding DWS, as shown in Figure 5:

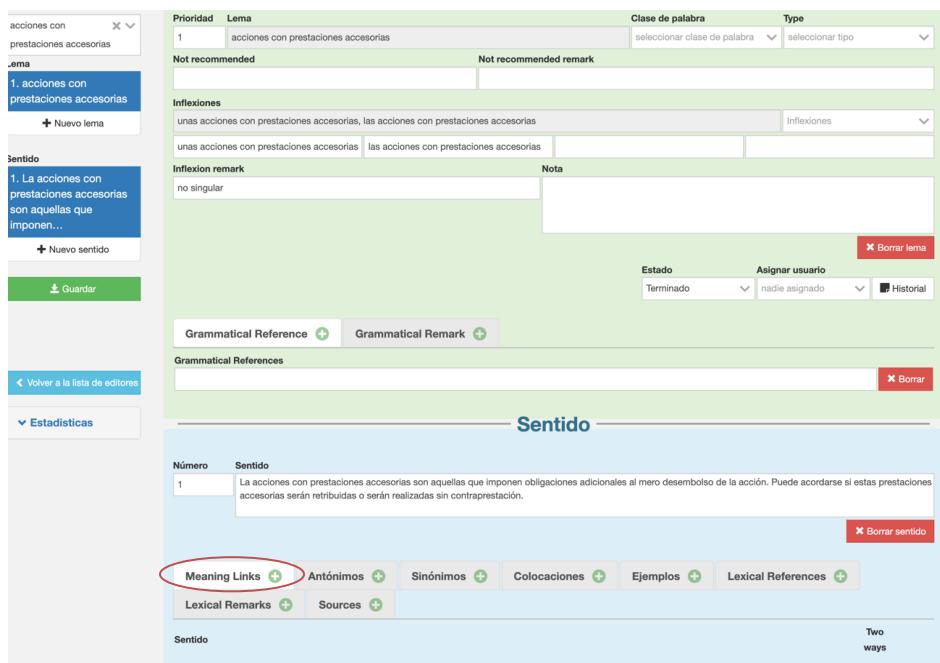


Figure 5: Automatic creation of the dictionary article for the accounting lemma **acciones con prestaciones accesorias** stored in the DWS of the accounting dictionary

By clicking on the button "Meaning Links" (circled in red in Figure 5), the system opens a window for writing the corresponding English lemma. If the English lemma is in the database, it will pop up and the other part of the dictionary will be automatically completed (Figure 6):

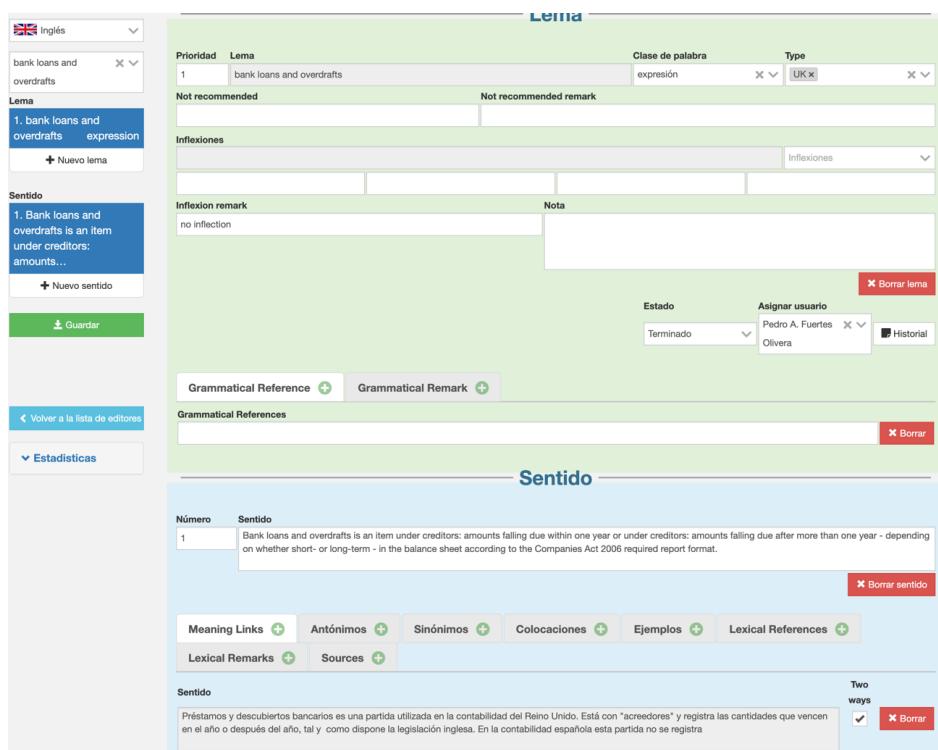


Figure 6: Automatic creation of the English lemma **bank loans and overdrafts**, with an English and Spanish definition

The two definitions always start with the lemma, followed by a verb. This makes very easy the process of searching in the tool and creating the dictionary articles automatically.

Finally, generative AI chatbots such as ChatGPT can be used for performing several lexicographic activities (see De Schryver 2023; Huete-García and Tarp 2024; Tarp and Nomdedeu-Rull 2024 for a critical analysis of the use of generative AI in lexicography). Figure 7 shows the conversation with the chatbot, initiated with an initial prompt, adapted from De Schryver (2023), about [PACAY]:

- PROMPT: Please give me lexicographic data for '[PACAY]'. Each sense should be in a numbered block. Each block then starts with the part of speech and the morphological forms of the respective sense. This is followed by a sense definition and sense examples that illustrate both the use and the meaning of each particular sense. For the example sentences, make sure to use different sentence structures, referring to different people; refer to past, present, and future situations; vary long and short example sentences; and include other elaborations, e.g. give me synonyms and countries in the Spanish speaking world where this word is used.

ChatGPT

 ChatGPT
[PACAY]

1. Sustantivo - Singular y Plural: Pacay / Adjetivo Relacionado: Pacayal

- *Definición:* Fruto de un árbol tropical perteneciente a la familia de las leguminosas (*Inga feuilleei*), con pulpa comestible de sabor dulce y textura suave.
- *Ejemplos:*
 - *En el mercado, compré una bolsa llena de pacayes maduros y jugosos.*
 - *Los niños disfrutaron de un refrescante pacayal durante la tarde caliente.*
- *Sinónimos:* guava, guama
- *Antónimos:* -

2. Sustantivo - Botánica: Pacay

- *Definición:* Árbol de la especie *Inga feuilleei*, cuyos frutos son los pacayes.
- *Ejemplos:*
 - *El pacay destaca en la selva amazónica por su altura y frondosidad.*
 - *Los agricultores plantaron pacayes en el huerto para cosechar los frutos.*
- *Sinónimos:* inga, guaba
- *Antónimos:* -

3. Sustantivo - Perú: Pacay



PF

PROMPT: Please tell me in which American countries this word is used because, to the best of my knowledge, PACAY is not used in Spain.

ChatGPT:

- La palabra "pacay" y sus significados son más comunes en países de América Latina, especialmente en regiones tropicales y subtropicales. Aunque las acepciones específicas pueden variar, la palabra es generalmente reconocida en contextos relacionados con frutas tropicales, árboles y bebidas tradicionales. Aquí hay algunos países de América donde se utiliza la palabra "pacay":
 1. Perú: En Perú, además de referirse al fruto y al árbol, "pacay" se utiliza específicamente para denominar una bebida tradicional fermentada hecha a partir del jugo de pacay.
 2. Ecuador: La palabra "pacay" es comúnmente utilizada en Ecuador para referirse al fruto y al árbol.
 3. Colombia: En algunas regiones de Colombia, como en la Amazonía, es probable encontrar el uso de la palabra "pacay" para hablar de esta fruta tropical.
 4. Venezuela: La palabra también se utiliza en Venezuela para referirse al árbol y al fruto.
- Es cierto que la palabra "pacay" y sus acepciones no son comunes en España, donde el léxico y las costumbres relacionadas con frutas tropicales pueden diferir. La variación en el uso de palabras es común en español debido a las diferencias culturales y geográficas entre las distintas regiones hispanohablantes.

Figure 7: Dialogue with ChatGPT on the Spanish word **pacay**

The above dialogue shows that working with chatbots such as ChatGPT has advantages and disadvantages. The former is that it can increase productivity, reduce lexicographic costs in both time and money, and allow searching for data that can be difficult to obtain, e.g., a particular meaning of a word which may be only used in one of the countries where Spanish is spoken (Spanish is spoken by more than 500 million native speakers). In future, it will be necessary to refine the practice of making prompts asking the chatbot for such data. Disadvantages are also well-known (see De Schryver 2023, Rundell 2023 and Huete-Garcia and Tarp 2024): hallucinations may be widespread and therefore, it is advisable to double check the data obtained with a chatbot before using them. These potential disadvantages, however, cannot make us forget the usefulness of Chatbots, e.g., example 2 was crafted with data shown in Figure 7.

5. Conclusion

This article has discussed the concept of sustainable lexicography in a rather different fashion to the one initially published by Colman (2016). The approach used here attempted to show that we must go beyond the language-centered lexicographic tradition that dominates current thinking and focus instead on new thinking centered on increasing lexicographic productivity and using technologies that (a) adopt a broad concept of lexicographic data, (b) speed up the lexicographic process, (c) save time and reduce costs, (d) facilitate direct cognitive processing, e.g. by machines, and (e) allow the individualization of data as units of consumption and sale. In particular, we must critically examine the benefits and drawbacks of the different practices on offer. The use of chatbots and other AI functionalities merit our consideration. I have no doubt that these will improve in time and that some of the qualms expressed these days by well-known scholars such as Vossen, (2022), Chomsky et al. (2023), McKean and Fitzgerald (2023) and Rundell (2023) will fade away.

Acknowledgment

Thanks are due to the participants in the symposium StellenLex 2024, held at Bureau of the Woordeboek van die Afrikaanse Taal, specially to the organizers Professors Rufus H. Gouws, Theo D. Bothma and Sven Tarp. I also thank the reviewers of this article and the editor of *Lexikos*, André du Plessis, for their comments on a previous draft of this paper.

Bibliography

Bergenholtz, H., S. Nielsen and S. Tarp (Eds.). 2009. *Lexicography at a Crossroads*. Bern: Peter Lang.
ChatGPT: <https://chat.openai.com/> (Access: February, 2024)

Chomsky, N., I. Roberts and J. Watumull. 2023. The False Promise of ChatGPT. *The New York Times*, 8 March 2023.
<https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html> (Access: April, 2024)

Colman, L. 2016. Sustainable Lexicography: Where to Go from Here with the ANW (*Algemeen Nederlands Woordenboek*, an Online General Language Dictionary of Contemporary Dutch?) *International Journal of Lexicography* 29(2): 139-155.
<https://doi.org/10.1093/ijl/ecw008>

DeepL Translate: <https://www.deepl.com/es/translator> (Access: April, 2024)

De Schryver, G.-M. 2023. Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography* 36(4): 355-387.
<https://doi.org/10.1093/ijl/ecad021>

DIDES. *Diccionario Digital del Español.* <https://diesgital.com/> (Access: April, 2024)

DLE. *Diccionario de la Lengua Española.* RAE. <https://dle.rae.es/> (Access: April, 2024)

Fuertes-Olivera, Pedro A. (Ed.). 2018. *The Routledge Handbook of Lexicography*. London/New York: Routledge.
<https://doi.org/10.4324/9781315104942>

Fuertes-Olivera, Pedro A. 2019. Designing and Making Commercially Driven Integrated Dictionary Portals: The *Diccionarios Valladolid-UVa. Lexicography* 6: 21-41.
<https://doi.org/10.1007/s40607-019-00056-8>

Fuertes-Olivera, Pedro A. 2022a. The Mental Lexicon in Lexicography: *The Diccionarios Valladolid-UVa. Lexikos* 32(1): 118-140.
DOI: <https://doi.org/10.5788/32-1-1712>

Fuertes-Olivera, Pedro A. 2022b. Theoretical, Technological and Financial Challenges: Some Reflections for Making Online Dictionaries. Jackson, Howard (Ed.). 2022. *The Bloomsbury Handbook of Lexicography*: 361-374. London/New Delhi/New York/Sydney: Bloomsbury Academic.
10.5040/9781350181731.ch-021

Fuertes-Olivera, Pedro A. and H. Bergenholz (Eds.). 2011. *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum.
10.5040/9781474211833

Fuertes-Olivera, Pedro A. and S. Tarp. 2014. *Theory and Practice of Specialised Online Dictionaries. Lexicography versus Terminography*. Berlin/Boston: De Gruyter.
<https://doi.org/10.1515/9783110349023>

Fuertes-Olivera, Pedro A. and S. Tarp. 2020. A Window to the Future: Proposal for a Lexico-graphically-assisted Writing Assistant. *Lexicographica* 36: 257-286.
<https://doi.org/10.1515/lex-2020-0014>

Fuertes-Olivera, Pedro A., S. Tarp and P. Sepstrup. 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., P. Gordo Gómez, M. Niño Amo, A. de los Ríos Rodicio, A. Sastre Ruano, S. Tarp, M. Velasco Sacristán, S. Nielsen, L. Mourier and H. Bergenholz. 2010. *Diccionario de Contabilidad Inglés-Español*. Navarra: Thomson Reuters-Aranzadi.

Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*. Oxford: OUP.
<https://doi.org/10.1093/acprof:oso/9780199654864.001.0001>

Huete-García, A. and S. Tarp. 2024. Training AI-based Writing Assistant for Spanish Learners: The Usefulness of Chatbots and the Indispensability of Human-assisted Intelligence. *Lexikos* 34(1): 21-40.
<https://doi.org/10.5788/34-1-1862>

Kosem, I., S. Krek and P. Gantar. 2021. Semantic Data Should no Longer Exist in Isolation: The Digital Dictionary Database of Slovenian. *Euralex 2020. Lexicography for Inclusion, 7–11 September 2021, Virtual*.
<https://elex.is/euralex2020/> (Access: April, 2024)

McKean, E. and W. Fitzgerald. 2023. The ROI of AI in Lexicography. *Proceedings of the 16th International Conference of the Asian Association for Lexicography: Lexicography (Asialex 2023 Proceedings)*, 22–24 June 2023, Seoul, Korea: Artificial Intelligence, and Dictionary Users: 10-20. Seoul: Yonsei University.

Moerdijk, F., C. Tiberius and J. Niestadt. 2008. Accessing the ANW Dictionary. Michael Zock and Chu-Ren Huang (Eds.). 2008. *Coling 2008: Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*, Manchester, 24 August 2008: 18-24. Manchester, UK: Coling 2008 Organizing Committee.
<https://aclanthology.org/W08-1900/>

Murphy, M.L. 2013. What We Talk about When We Talk about Synonyms (And What It Can Tell Us about Thesauruses). *International Journal of Lexicography* 26(3): 279-304.

Nomdedeu-Rull, A. and S. Tarp. 2024. *Introducción a la lexicografía en español. Funciones y aplicaciones*. London: Routledge.

Rundell, M. 2023. Automating the Creation of Dictionaries: Are We Nearly There? *Proceedings of the 16th International Conference of the Asian Association for Lexicography: Lexicography (Asialex 2023 Proceedings)*, 22–24 June 2023, Seoul, Korea: Artificial Intelligence, and Dictionary Users: 1-9. Seoul: Yonsei University.

Tarp, S. 2011. Lexicographic and Other e-Tools for Consultation Purposes: Towards the Individualization of Needs Satisfaction. Fuertes-Olivera, Pedro A. and Henning Bergenholz (Eds.). 2011. *e-Lexicography: The Internet, Digital Initiatives and Lexicography*: 54-70. London/New York: Continuum.

Tarp, S. 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, S. and Pedro A. Fuertes-Olivera. 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>

Tarp, S. and A. Nomdedeu-Rull. 2024. Who Has the Last Word? Lessons from Using ChatGPT to Develop an AI-based Spanish Writing Assistant. *Círculo de lingüística aplicada a la comunicación* 97: 309-321.
<https://dx.doi.org/10.5209/clac.91985>

Tiberius, C., J. Kallas, S. Koeva, M. Langemets and I. Kosem. 2024. A Lexicographic Practice Map of Europe. *International Journal of Lexicography* 37(1): 1-28.
<https://doi.org/10.1093/ijl/ecad023>

Vossen, P. 2022. ChatGPT Is a Waste of Time. *VU Magazine*, 22 December 2022.
<https://vumagazine.nl/professor-piek-vossen-chatgpt-is-a-waste-of-time?lang=e>