

Turning Bilingual Lexicography Upside Down: Improving Quality and Productivity with New Methods and Technology

Sven Tarp, *Centre for Lexicographical Studies, Guangdong University of Foreign Studies, China; Department of Afrikaans and Dutch, Stellenbosch University, South Africa; International Centre for Lexicography, Valladolid University, Spain; Centre of Excellence in Language Technology, Ordbogen A/S, Denmark; and Centre for Lexicography, Aarhus University, Denmark (st@cc.au.dk)*

Abstract: This is a report from the real world. It informs about the outcome of a project, which the author conducted during a months-long research stay at the Danish company Ordbogen where he integrated its research and development (R&D) team. The first part of the project was to test machine translation and find out to what extent it is usable in the compilation of bilingual lexicographical databases. The hypothesis was that the technology was not yet mature. But surprisingly, it turned out that the accuracy rate is already so high that it is worth considering how to implement it. The second part of the project aimed at further developing an idea formulated by Fuertes-Olivera et al. (2018) on how to invert a dictionary without losing semantic content. The new vision is to compile a monolingual L2 database, bilingualize it to an L2–L1 database using machine translation, and then invert the relationship between L2 lemmata and L1 equivalents using the L1 definitions of the L2 lemmata as the axis. The third part of the project was to test this idea using a specially designed ad hoc program. The program automatically uploads relevant data from existing lexicographical databases, translates L2 definitions and example sentences into L1, suggests adequate L1 equivalents, and eventually inverts the relationship between the two languages. It worked, but the methodology still needs further refinement to be implementable on a large scale. The report concludes by listing some of the remaining challenges and defining the new role of the lexicographer in this type of project.

Keywords: LEXICOGRAPHICAL R&D, INTERDISCIPLINARY COLLABORATION, DIGITAL TECHNOLOGY, BILINGUAL LEXICOGRAPHY, LEXICOGRAPHICAL DATABASES, MACHINE TRANSLATION, AUTOMATIC INVERSION, OBJECT LANGUAGE, AUXILIARY LANGUAGE, HUMAN VERSUS ARTIFICIAL LEXICOGRAPHER

Opsomming: 'n Omwenteling in tweetalige leksikografie: Die verbetering van kwaliteit en produktiwiteit met nuwe metodes en tegnologie. Hierdie is 'n verslag uit die praktyk. Daar word verslag gelewer oor die uitkoms van 'n projek, wat die outeur gedurende 'n maandelange navorsingstydperk by die Deense maatskappy Ordbogen aangevoer het waartydens hy die navorsings- en ontwikkeling- (N&O)-span geïntegreer het. Die eerste deel van

die projek het die toets van masjienvertaling behels en om die bruikbaarheid daarvan in die saamstel van tweetalige leksikografiese databasisse te bepaal. Die hipotese was dat die tegnologie nog nie gevorderd genoeg was nie. Dit het egter, verrassend genoeg, geblyk dat die akkuraatheidsyfer reeds so hoog was dat dit die moeite werd was om die implementering daarvan te oorweeg. Die tweede deel van die projek het die verdere ontwikkeling van 'n idee, geformuleer deur Fuertes-Olivera et al. (2018) oor die omskakeling van 'n woordeboek sonder verlies van semantiese inhoud, ten doel gehad. Die nuwe visie is om 'n eentalige L2-databasis saam te stel, dan met behulp van masjienvertaling te omskep in 'n L2-L1-databasis, en daarna die verhouding tussen L2-lemmata en L1-ekwivalente om te skakel deur die L1-definisies van die L2-lemmata as die spil te gebruik. Die derde deel van die projek was die toets van hierdie idee met 'n spesiaal ontwerpte ad hoc-program. Hierdie program laai outomaties relevante data vanuit leksikografiese databasisse, vertaal L2-definisies en -voorbeeldsinne in L1, stel gepaste L1-ekwivalente voor, en skakel uiteindelik die verhouding tussen die twee tale om. Dit was geslaagd, maar die metodologie moet nog verder verfyn word voordat dit op groot skaal geïmplementeer kan word. Die verslag word afgesluit met die lys van sommige van die oorblywende uitdagings en met die definiëring van die nuwe rol van die leksikograaf in hierdie tipe projek.

Slutelwoorde: LEKSIKOGRAFIESE N&O, INTERDISCIPLINÊRE SAMEWERKING, DIGITALE TEGNOLOGIE, TWEETALIGE LEKSIKOGRAFIE, LEKSIKOGRAFIESE DATABASISSE, MASJIENVERTALING, OUTOMATIESE OMSKAKELING, DOELTAAL, SEKONDÊRE TAAL, MENSLIKE VERSUS KUNSMATIGE LEKSIKOGRAAF

1. Introduction

Today, the compilation and presentation of dictionaries and other lexicographical products are inconceivable without assistance from digital technologies that are constantly improving and breaking new ground. Good lexicographical craft presupposes, in one way or another, interdisciplinary collaboration with specialists from other fields. The collaboration has various dimensions. One of them is research and development (R&D), i.e., experimenting with new compilation methods and ways of presenting the final product to the target users. To be successful, the pursuit of innovation requires almost daily contact and daily exchange of views between the lexicographer, on the one hand, and information scientists, programmers, or designers, on the other hand. Few lexicographers, whether university professors or employees in publishing houses, have the opportunity to engage in this kind of interdisciplinary research, without which the discipline will have to struggle even more to find its place in the Fourth Industrial Revolution. Far too often, the two parts work in different directions with only occasional contact.

From this perspective, I consider myself extremely privileged. In 2021, my University facilitated a research stay at Ordbogen A/S, a successful Danish company specializing in language services, digital teaching material, online dictionaries, and writing assistants. It was an extraordinary experience. It allowed me to conduct experiments and test new and old ideas during various

months. From the very first day, I was co-opted by the company's ODIN Team, where most of its research and development takes place. Here, I was the only one with a background in lexicography, language didactics, and translation. The other members were information scientists, programmers, and designers who, from the perspective of their expertise, had a very different approach to my discipline. It was both challenging and stimulating.

Since 2017, I have, to some extent, collaborated in the development of Ordbogen's digital Write Assistant and published several research articles on this topic, one of them together with information scientists from the company (Tarp et al. 2017). I find the underlying philosophy and technology timely and future-oriented. I am especially fascinated by the new ways and forms in which lexicographical data can be presented to users engaged in a particular activity like L2 writing, thus supporting the basic tenets of the Function Theory (see Fuertes-Olivera and Tarp 2014). But I am also very critical of how it has been done so far. From my narrow disciplinary perspective, I have even described the lexicographical component as "Write Assistant's Achilles heel", among other things because "existing databases are highly deficient and problematic" when it comes to serving this kind of software and turning it into a high-quality product (Tarp 2019: 237-238).

There is an urgent need to prepare more appropriate databases, and to this end, develop new compilation methods and techniques that can guarantee both higher quality and higher productivity. The last point is particularly important considering that many publishers of dictionaries struggle to make ends meet. Higher productivity without compromising quality could be part of the solution, but it calls for basic research through intense interdisciplinary collaboration between relevant experts. I was thus excited when Aarhus University allowed me to have a research stay at Ordbogen. To that end, I formulated two projects, or experiments, to be conducted:

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 function as semantic differentiators when bridging from L1 to L2.

The immediate objective of the two experiments was to see how far technology has come, to what extent it is already implementable, and what consequences it may have for the future relationship between man and machine, between the human and the artificial lexicographer. The following report and reflections will focus on the second of the two experiments, which resulted in a major breakthrough.

2. Testing machine translation

After decades of struggle, machine translation has improved considerably

during the past few years. I am not aware of any lexicographers experimenting with this technology, but I am convinced that it is only a question of time before the discipline will adopt it at a broad scale. It is thus necessary to be at the forefront of this development, although I did not have high expectations for its immediate relevance to lexicography. At our initial meeting, Michael Walther, head of the ODIN Team, said that this technology only becomes really attractive and relevant when over 70 percent of the translations are correct and can be inserted directly into the database. Both he and I judged that it still takes some time before we can achieve this success rate.

During the whole research stay, I worked closely with Henrik Hoffmann, a talented programmer and web developer. We decided to start with Google Translate, which many scholars consider the best, or at least one of the best, translation tools available today. As an empirical basis, we chose a monolingual Spanish database which Pedro Fuertes-Olivera is compiling in Valladolid under the auspices of Ordbogen (at that moment, he had completed approximately 80,000 senses). Henrik Hoffmann then extracted 200 random definitions, which were immediately translated into English using Google Translate.

It was then my task to systematically compare source and target definitions. The result was disappointing, but as expected. About 30 percent of the translations were acceptable, though not perfect. Another 40 percent had major or minor errors that were disruptive to understanding. And the last 30 percent were straight-out incomprehensible.

At this point, it is pertinent to make some observations:

1. Translating 200 definitions out of 80,000 does not give statistically precise results, but the tendency is convincing, according to discussions with members of the ODIN Team;
2. The definitions are written in the non-natural language that characterizes most dictionary definitions. This phenomenon may present additional challenges for machine translation;
3. Spanish is generally a synthetic language, a characteristic that may give rise to particular types of problems when translating into a predominantly analytical language such as English;
4. The result cannot be generalized and directly applied to translation in the reverse language direction, i.e., from English to Spanish, or to translation between other language pairs, each of which has its characteristic features.

Either way, Google Translate is not the solution to our problem. Henrik Hoffmann therefore suggested that we test the DeepL Translator, which I had not used before. The results of this new test came as a complete surprise. Of 200 translated definitions (the same as previously), 156 (78%) were now completely correct, while 44 (22%) had major or minor errors. Most surprisingly, DeepL

correctly translated most cases where Google Translate had to give up on the particular syntactic structure of the Spanish definitions.

Valladolid-UVa	Google Translate	DeepL Translator
abacorar VERB producir una sensación de angustia o recelo	produce a feeling of anxiety or apprehension	to produce a feeling of distress or apprehension
abajar VERB mover algo o moverse uno mismo de un lugar o posición superior a un lugar o posición inferior	move something or move one of a place or a place higher position or lower position itself	to move something or to move oneself from a higher place or position to a lower place or position
índice NOUN en imprenta, símbolo tipográfico en forma de mano que apunta un lugar; se usa para señalar algo importante o para llamar la atención en un texto	in press, typographic symbol shaped hand pointing a location; It is used to indicate important something or to get attention in a text	in print, typographic symbol in the form of a hand pointing to a place; used to point to something important or to draw attention to a text
ígneo ADJECTIVE que hace referencia a las rocas, minerales o formaciones geológicas que se originaron a partir de los materiales fundidos del interior de la tierra	It referred to rocks, minerals or geologic formations that originated from molten material of the inner earth	which refers to rocks, minerals or geological formations that originated from the molten materials of the earth's interior

Table 1: Comparison of Google's and DeepL's translation performance

Table 1 contains four illustrative examples of how Google Translate and DeepL perform when translating definitions from the Spanish database (Valladolid-UVa). The definition assigned to the verb *abacorar* has a relatively simple structure, and both translation tools provide English translations that are correct and understandable, although slightly different. By contrast, the definitions of the verb *abajar* and the noun *índice* have a more complex syntactic structure. In both cases, Google Translate offers incomprehensible translations, whereas DeepL manages to decipher the Spanish syntax and provide acceptable translations into English. Finally, the adjective *ígneo* shows a minor but frequent problem where Google Translate, contrary to DeepL, cannot grasp the initial structure of the Spanish definition and provides an English text that is understandable but requires language revision to serve as a lexicographical definition.

Just in case, we had Pedro Fuertes-Olivera review the translations, and he reached the same result. In other words, the performance of DeepL looked more than promising for the use of machine translation in lexicography, but further testing was required to give a final judgment.

The target users of the definitions contained in the monolingual Spanish database are native Spanish speakers. However, since the translation of these items only makes sense if the target users are non-native speakers, we now decided to extract our empirical data from a monolingual dictionary designed for the latter. For this purpose, we chose the *Oxford Dictionary* on the Lexico platform, from which we hand-picked 200 random definitions belonging to different letters and parts of speech. These definitions are characterized by a more straightforward language with less complex syntax than those from the Valladolid-UVa database. We now used DeepL to translate them into Spanish and Danish, respectively. The result was even more convincing than previously.

Of 200 definitions translated into Spanish, 187 (93,5%) were now completely correct, while 13 (6,5%) had minor or major problems. Of these, only 4 (2%) were so serious as to require a total rewrite, whereas the rest only needed a slight reworking. The revision could, in most cases, be done using another of DeepL's functionalities (see below).

Spanish and English are languages with many native speakers, while Danish has a lot fewer. As DeepL is trained on existing texts and translations, we expected its performance to be somewhat inferior in Danish. This prediction turned out to be true, but the result was far from catastrophic. Of 200 definitions translated into Danish, 142 (71%) were correct, 48 (24%) had minor problems, while the remaining 10 (5%) were unusable. In half of the 48 translations with minor problems, the first word in the definition should either be deleted or changed from one word class or inflectional form to another. Another frequent problem was the appearance of two identical words in the translation when the source definition included two more or less synonymous words. In both cases, these are minor inconveniences, which the lexicographers revising the text can quickly and easily correct by clicking on the alternative suggestions that DeepL offers its users.

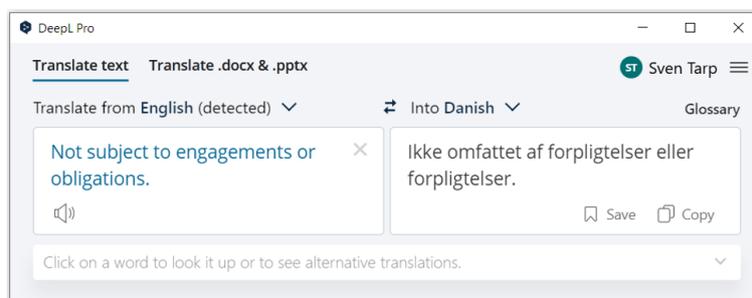


Figure 1: Using DeepL to translate from English into Danish

In Figures 1 and 2 we can see how it works. Figure 1 shows the translation of the English text segment *Not subject to engagements or obligations* that defines one of the senses of the adjective *free*. The words *engagements* and *obligations* are synonymous to some extent. However, DeepL does not grasp the subtle semantic difference and translates both words into *forpligtelser*. The Danish definition is understandable. But it is not as semantically rich as the English one, and neither is it stylistically convenient in a lexicographical context. To solve this problem, the user of DeepL can simply click on one of the translated words and get alternative solutions (see Figure 2). In the concrete case, at least two suggested alternatives (*engagementer* and *aftaler*) could replace *forpligtelser* the first time it occurs. Another click on the preferred suggestion modifies the target definition accordingly. As can be seen, DeepL's user-friendly design and functionalities make it easy to revise and, if needed, correct small pieces of text. A trained lexicographer can probably do it in a few seconds, thus saving a considerable amount of time.

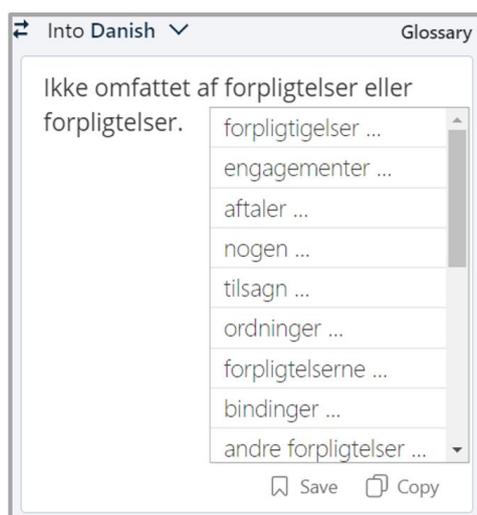


Figure 2: Alternative solutions suggested by DeepL

As mentioned above, our findings are not statistically precise, as the exact percentages cannot be repeated. But the tendency is indisputable. Out of curiosity, we also had colleagues revise a few samples of the same English definitions translated into Chinese, French, and German, respectively. In all three cases, DeepL showed a very high accuracy rate. The conclusion was that we were, indeed, on the verge of a breakthrough that may have important implications for the compilation of future bilingual lexicographical databases. But what are the implications? It was now time to take advantage of this unique opportunity

to experiment with new methods for generating lexicographical data. It implied, above all, to elaborate on an old idea.

3. Reflections on bilingual lexicography

For several years, I have been critical of the way bilingual dictionaries and lexicographic databases are conceived. The very concept of a bilingual dictionary is clearly ambiguous (see Tarp 2005). As used in both academic and non-academic literature, it covers a broad range of very different dictionary types, where the only common feature seems to be that two languages are involved in one way the other. The definitions provided by well-known dictionaries of lexicography like Martínez de Sousa (1995), Bergenholtz et al. (1997), Burkhanov (1998), and Hartmann and James (1998) also vary considerably. Some scholars even regard the terms *bilingual dictionary* and *translation dictionary* as synonyms. Marelló (2003: 325), for instance, defines "bilingual dictionaries only as those dictionaries which place the two languages in contact for purposes of translation". Several terms imported from translation science have strongly influenced bilingual lexicographical terminology. In his classical book on new French dictionaries, Hausmann (1977: 58) used the terms *herübersetzende* and *hinübersetzende* (*translating into* and *from L1*, respectively) to classify L2–L1 and L1–L2 dictionaries.

I have become increasingly convinced that existing lexicographical terminology mentally blocks the necessary creativity and innovation. This is particularly true of the terms *source language* and *target language*, which also have been adopted from translation studies. They are applied almost uncritically to denote L1 and L2 in an L1–L2 dictionary and L2 and L1 in an L2–L1 dictionary, respectively, i.e., sometimes referring to one language and sometimes to the other.

Fuertes-Olivera et al. (2018) discussed the negative consequences of this terminology in the context of bilingual dictionary compilation and, instead, introduced a new vision based on the alternative terms *object language* and *auxiliary language*, where the latter is always the user's native language and the former a foreign, or non-native, language, regardless of the language direction of a dictionary or database.

The rationale is the simple one that an L1 user, as a rule, looks up in an L2–L1 dictionary to understand or know something about L2, while the same user looks up in an L1–L2 dictionary to find an L2 word or L2 expression. In both cases, the lookup is about L2, which is thus the object of the lookup. By contrast, the function of L1 is to help the user either understand L2, use L2, or bridge to L2. The practical implications are considerable. Bilingual L1–L2 dictionaries have so far been built on a selection of L1 lemmata. Instead, the new vision is to base them on a stock of L2 lemmata, which are then processed into L2–L1 and subsequently inverted into L1–L2 dictionaries. In the following, I will briefly explain how this idea has developed and what the implications are.

3.1 Scerba's translating dictionaries

Few people accomplish two or more languages to the same perfection. Most people have a first language (mother tongue or native language), which they master far better than other languages and may, therefore, be considered learners of these languages, whatever their proficiency level. As such, they may need dictionaries when they study and communicate in them. From the perspective of lexicography, the question has been which type of dictionary will best serve this purpose: a monolingual or a bilingual one? This discussion has been going on for decades.

More than eighty years ago, Scerba (1940: 341) urged L2 learners "to discard translating dictionaries as soon as possible and switch to the defining dictionary of the foreign language." By *translating dictionaries*, the Russian scholar understood dictionaries with their lemmata translated into equivalents in another language. Scerba had a distinct language-didactic approach and opposed the contrastive method used in language teaching because it could lead to "a mixed bilingualism due to numerous transfers from L1", and consequently "only goes for L2 learning at the beginner's level" (Mikkelsen 1992: 34). In Scerba's opinion, the L1–L2 "translating dictionary" was only beneficial for L2 learners at a beginner's level. To serve this user segment, he defined a set of general principles for "a special type of translation dictionary" from L1 to L2 (see Mikkelsen 1992: 27). On the other hand, he proposed to write L1 definitions in the L2 "defining dictionaries":

One could create foreign defining dictionaries in the students' native language. Of course, translations of words could also be included when this would simplify definition and would not be detrimental to a full understanding of the foreign word's true nature. (Scerba 1940: 341)

The suggestions show that Scerba was aware of the L2 learners' need to be assisted in their mother tongue, at least for a period. But he regarded it as a *malum necessarium* and advised the users to switch to monolingual L2 dictionaries "as soon as possible". For many years, I was captivated by Scerba's ideas. Today, I am more reserved and hold his disciplinary approach to be somehow misguided inasmuch as a learner's dictionary is not a learning tool in itself but a consultation tool designed to provide the best possible assistance to the L2-learning process. Scerba's ideas were generally unknown in the West until the late 20th Century, but they were applied with minor adjustments in the former Soviet Union and part of Eastern Europe. The results were apparently positive, though not unequivocally. Duda (1986), for instance, reported various critical points appearing in reviews of a Russian–German defining dictionary that had applied Scerba's principles. Among the criticisms were the user's difficulties to perform lexicalization in their native language:

The user is apparently able to understand the analysis of a word's meaning as it is given in the definition in a monolingual dictionary. It appears to be much

more difficult for the user to state the meaning based on a given definition, i.e. to perform lexicalization. (Duda 1986: 13)

Duda did not find that the suggestion substantially challenged Scerba's concept for a defining dictionary. But it required that his principle of providing definitions rather than equivalents was changed so that the latter had priority whenever possible and adequate. This adjustment was definitely a case for bilingual learners' dictionaries.

3.2 The Big Five

In the West, a different tradition developed. During various decades, monolingual dictionaries were considered the *crème de la crème* of learners' dictionaries, notably the English Big Five (Oxford, Longman, Collins, Cambridge, and MacMillan), which largely influenced Western lexicographical thinking after English became an international lingua franca. These dictionaries were undoubtedly high-standard with many innovative and sophisticated features. But just like Hollywood stars, they were iconized and turned into a sort of one-size-fits-all product for users with very different linguistic and cultural backgrounds. Their advocates followed the philosophy that dominated language didactics for many years. Learners should be forced into thinking in the foreign language and, therefore, not unnecessarily exposed to their mother tongue during the learning process. From this perspective, dictionaries were but one of many learning tools and should, by definition, be monolingual. But many language teachers also observed that only a minority of students followed their recommendation, at least outside the classroom where they consulted all sorts of bilingual dictionaries to get the required assistance. The central role of the native language for most L2 learners was clearly underestimated. In this respect, Adamska-Sałaciak and Kernerman (2016: 273) rightly state that the monolingual learners' dictionary "focuses practically on an elite minority of top-level users." Lexicographers have to act accordingly.

Over the years, the critical voices became louder and louder, thus making a new case for bilingual learners' dictionaries (see Tomaszczyk 1983, Lew and Adamska-Sałaciak 2015, among others). At first, we saw the so-called semi-bilingual dictionaries that were influenced by the then predominant language-didactic philosophy. They consisted of "traditional" L2 monolingual dictionaries that were bilingualized with the addition of equivalents in the learners' native language. The idea was to force the users to read the L2 definitions and, only in case they had problems, resort to the equivalents. It did not really work, so little by little came more "pure" bilingual L2–L1 dictionaries without definitions neither in L1 nor L2, i.e., only with L1 equivalents.

Generally, the renewal progressed without sufficiently taking account of the new digital technologies. On the one hand, the supporters of a paradigm shift acknowledged the high-quality features of monolingual dictionaries like

the Big Five and wanted to incorporate them in a new type of bilingual dictionary that would let "users have the best of both worlds" (Adamska-Sałaciak 2010: 123). On the other hand, many new dictionary projects seemed stuck in the old dichotomy between monolingual and bilingual. In an article on the recent development of learners' dictionaries, I argued that:

(...) the formal classification of dictionaries for foreign-language learners into monolingual and bilingual may be one of the major obstacles that prevent present-day lexicography from taking full advantage of the new technologies and designing the dictionaries that meet the real needs of foreign-language learners. (Tarp 2013: 426)

The introduction of digital technologies in lexicography is disruptive. It requires that the discipline be reconsidered from top to bottom, especially the compilation, storing, and presentation of lexicographical data. The fact that the results of the lexicographers' efforts are stored in databases implies that these results can be presented to the target users in differentiated and needs-adapted ways and quantities through the method of filtering (see Bothma 2011). A digital dictionary — that is, the set of lexicographical data visualized in user interfaces — can now elegantly combine features of traditional monolingual and bilingual dictionaries. The lexicographical data stored in the database and those uploaded in user interfaces are not necessarily identical. Some data can even be stored mainly or exclusively for compilation purposes.

3.3 The Valladolid experience

I mentioned previously that the replacement of the terms *source* and *target language* with *object* and *auxiliary language* has considerable practical implications. Whereas bilingual L2–L1 dictionaries, as a rule, are either bilingualized extensions of existing L2 dictionaries or built on an independently selected stock of L2 lemmata, bilingual L1–L2 dictionaries have so far been compiled from a selection of L1 lemmata. There are very few exceptions to this. One of these is a mono-directional, biscopal English–Spanish dictionary project, which Pedro Fuertes-Olivera is working on in Valladolid. Its target users are Spanish native speakers, which means that the object language is English and the auxiliary one Spanish. This vision informed the decision to apply an innovative compilation methodology in the project, as summarized by Fuertes-Olivera et al. (2018: 160):

Whereas traditional mono-directional, biscopal dictionary projects usually take their point of departure in the users' native language, the Valladolid project does the opposite. It starts with a selection and description of English lemmata including separation in senses, definitions, Spanish equivalents, grammar, etc. An automatic and simultaneous inversion is then made where the Spanish equivalents to one of more English lemmata become new lemmata whereas the English lemmata become equivalents with the brief Spanish definitions used as meaning discrimination.

The real innovation here is the use of L2 definitions written in L1 as semantic differentiators in an L1–L2 dictionary. The inversion of dictionaries — or lexicographical databases, which is a more precise term — should be done without losing semantic content. Except for some specialized dictionaries within culture-independent disciplines, this challenge has cast a shadow over virtually all previous attempts of automatic dictionary inversion. The described compilation methodology, which Ordbogen's programmers incorporated into the Dictionary Writing System in close collaboration with the project editor, should ensure this. It worked! There were, of course, various new challenges — among them, the nature of the L2 definitions, the revision of the L1–L2 part, and the lexicographer's new role — which I will discuss and further elaborate on in the following.

3.4 What is to be done?

The main drive behind the methodology used in the Valladolid project was to raise productivity without compromising quality. I regard the described methodology as an important step forward in achieving this goal. But the experience from the machine-translation experiments indicates that there is still a job to do. I thus returned to an old idea that I discussed but left unfinished in my second doctoral thesis (Tarp 2008). The idea took form during the lecture of a large body of relevant literature (some of it mentioned above) and the study of hundreds of dictionaries, among which the Swedish LEXIN project targeting immigrants from more than 20 language communities was particularly inspiring (see Gellerstam 1999). But the idea could not be fully developed at that moment due to my insufficient knowledge of and experience with digital technologies.

I realized that the research stay at Ordbogen now gave me a unique chance to refine the idea and make it directly implementable. It goes without saying that it took a great deal of reflection to connect the dots and get the complete picture. I got the inspiration not by buying pastries for five cents down at the bakery, like Jack London, but on long evening walks in the countryside around Odense. All in all, I must have walked over 150 km in these beautiful surroundings to clear my thoughts. But it was worth it! The main idea, which I presented to Henrik Hoffmann, was as follows:

1. A monolingual L2 lexicographic database is compiled, intended for L1 users. This core database serves as a *basis* for the further process. It resembles Scerba's "defining dictionary" and incorporates, at the same time, many relevant features from the Big Five.
2. The core database is bilingualized into an L2–L1 database. The conversion is done by automatically translating definitions and example sentences (and metalanguage) and assigning L1 equivalents to the respective senses of each lemma.

3. The relationship between L2 lemmata and L1 equivalents contained in the database is inverted, so that the latter become lemmata and the former equivalents. The original and now translated definitions of L2 words — together with the corresponding L2 example sentences and their L1 translations — follow these words throughout, including during their metamorphosis from lemmata to equivalents. In this process, the definitions themselves slough their skin and transmute into differentiators. The new L1–L2 relationship is not a traditional one that contrasts two languages but a bridge from L1 into L2. This difference is attributable to the definition-turned-into-differentiator that directly states the meaning of the relevant sense of the L2 equivalent without focusing on the semantic differences between the treated sense of the L1 lemma and its L2 equivalent.
4. This methodology increases productivity because much of the process is mechanical. But it also improves the overall quality of the final product since L2 words — whether they appear as lemmata or equivalents — are never detached from their original semantic and syntactic L2 universe, as frequently happens in traditional L1–L2 dictionaries.
5. As a spin-off, the idea also facilitates economy of scale. Applying the described methodology, once a core database has been compiled in one language, say English, it is relatively easy to convert it into a set of bilingual dictionaries between this language and (many) other languages like Spanish, Danish, German, French, and Chinese.

4. Testing the idea

The discussion with Henrik Hoffman was difficult but highly productive. I presented the idea, he listened and had some suggestions which I opposed. He insisted that we reuse some of the lexicographical data in Ordbogen's databases, arguing that it would be more convincing. I rejected this idea because I was afraid it might jeopardize the quality of the core database and, consequently, its bilingual extensions. Both of us stood firm on our opinions, and the discussion went on for several days. Little by little, we began understanding each other's points of view and eventually reached a consensus. It was a beautiful interdisciplinary experience.

4.1 The first test

When Henrik Hoffman had grasped the idea, he immediately sat down and prepared an ad-hoc program in just two hours! To test it, he uploaded data from Politiken's English–Danish dictionary for advanced learners (*Store Engelskordbog*), which Ordbogen took over from the Politiken Publishing House after the latter's dictionary department closed down. The dictionary is a "semi-bilingual"

extension of *Collins Cobuild English Dictionary*, which includes Danish equivalents placed before (!) the English definitions, and has the metalanguage reproduced in Danish (see Figure 3). He chose this particular dictionary because it was easier to test, as it already contains Danish equivalents. When we typed an English lemma in the search field and clicked on the magnifier, the program automatically translated the definitions and example sentences using DeepL.

Figure 4 shows the result. It is not a dictionary article but only the visualization of some of the lexicographical data stored in the database. How it eventually will be presented to the users depends on the specific purpose, i.e., whether it is for a dictionary, an e-reader, or a writing assistant (see Fuertes-Olivera and Tarp 2020). The figure gives us a brief idea of the quality of the translated example sentences, which is, at least, as high as the quality of the translated definitions tested in Section 2. Contrary to the latter, these sentences represent natural language, so it is hardly a surprise. In this test, we were only interested in the program's functionality and immediate performance. We thus uploaded, studied, and saved a number of L2 lemmata with the attached data (definitions translated into L1, L1 equivalents, L2 example sentences, and their L2 translations). We then skipped the revision and correction phase and went directly to the Danish–English part, where we analyzed the data assigned to some of the inverted Danish lemmata. One of these was the reflexive verb *opføre sig*, which appeared twice as equivalent to *behave* in the English–Danish part (see Figure 5).

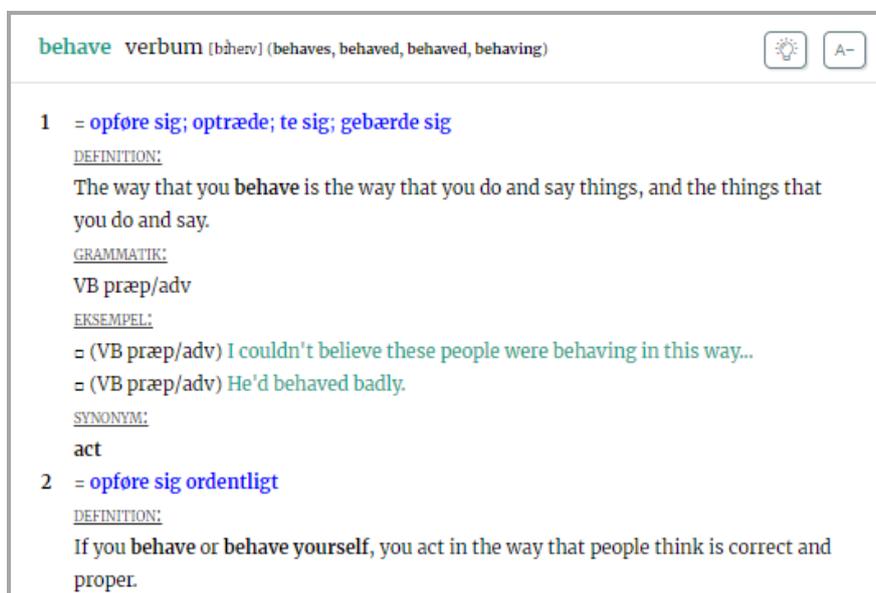


Figure 3: Extract of article from Politiken's English–Danish dictionary

behave 🔍 English Danish ▼

behave, verbum
behaves, behaved, behaved, behaving

1. Den måde, du opfører dig på, er den måde, du gør og siger ting på, og de ting, du gør og siger.

Ækvivalenter:

- opføre sig
- optræde
- te sig
- gebærde sig

Eksempler:

- I couldn't believe these people were behaving in this way... - *Jeg kunne ikke tro, at disse mennesker opførte sig på denne måde...*
- He'd behaved badly. - *Han havde opført sig dårligt.*

Figure 4: Translation of sense from Politiken's dictionary

opføre sig 🔍 Danish English ▼

opføre sig, verbum

1. (Den måde, du opfører dig på, er den måde, du gør og siger ting på, og de ting, du gør og siger.)
behave
Eksempler:

- I couldn't believe these people were behaving in this way... - *Jeg kunne ikke tro, at disse mennesker opførte sig på denne måde...*
- He'd behaved badly. - *Han havde opført sig dårligt.*

2. (Inden for videnskaben er den måde, noget opfører sig på, de ting, det gør.)
behave
Eksempler:

- Under certain conditions, electrons can behave like waves rather than particles. - *Under visse betingelser kan elektroner opføre sig som bølger i stedet for partikler.*

Figure 5: Inversion with ad-hoc program

We learned several things from this first test. First of all, that the overall idea does work but not surprisingly needs further refinement. It confirmed that special attention should be paid to the translated L2 definition, as it is the central axis around which the whole project rotates. In the test, we used Collins Cobuild's innovative "new definition" written in a natural language similar to the one that teachers use when explaining L2 words to their students. This defining technique implies that the defined word is part of the definition. One may like or dislike this type of definition (personally, I like it). But as can be seen in Figure 5, it does not work in our context because the translation defines the L1 word instead of the L2 one.

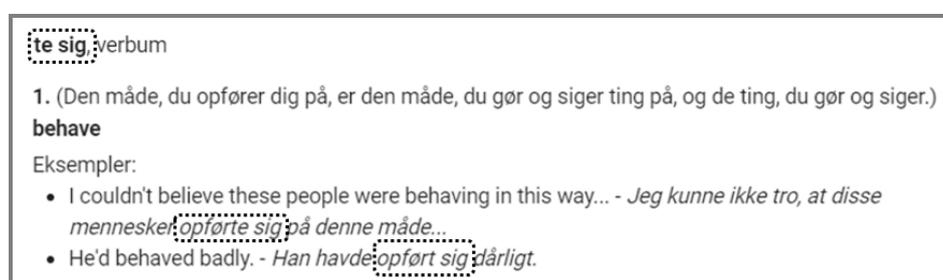


Figure 6: Discrepancy between lemma and translated example sentence

Another minor problem detected was the translation of example sentences when more than one equivalent were assigned to an L2 sense, for instance, the first sense of *behave* in Figure 4. In this case, the translations include the first of these equivalents (*opføre sig*). However, if one of the other equivalents appears as a lemma in the Danish–English part (e.g., *te sig*), there will be a certain discrepancy between the lemma and the translated example sentence (see Figure 6). It is not a big problem because the example sentence here aims at illustrating a specific syntactic property of *behave*, and the translation is merely provided to assist L1 users at a low L2 proficiency level who do not understand it. But overall it would be better to adjust it. Finally, we had no trouble admitting that we were lucky to find a learners' dictionary that provided both L2 definitions and L1 equivalents. Such dictionaries are rare, at least with the required quality. This takes us to the second test.

4.2 The second test

Henrik Hoffmann had now improved the ad-hoc program adding more functionalities. This time, we decided to upload a number of lemmata inclusive definitions from the *Oxford Dictionary of English*, which Ordbogen also licenses. The definitions were automatically translated with DeepL and could also be

corrected if there were any mistakes. The corrections, however, had to be done manually, as we did not yet have the license to build all DeepL's services into the software.

When the translations were accepted and saved, a new page appeared with two columns. The left one displayed the translated definitions, whereas the right one offered possible equivalents to the lemma. The program automatically retrieved these L1 candidates from Ordbogen's self-produced English–Danish database. Now, I finally understood the full implications of what Henrik meant when he said he wanted to reuse lexicographical data from the company's own databases. It really makes things easier. After carefully reading the definitions of the various senses of the lemma, the lexicographer can assign the suggested equivalents to the appropriate senses with a simple movement (see Figure 7) and, if necessary, add additional ones. If we had used some of the other English–Danish databases that the company manages, the number of suggested equivalents would probably be higher, thus facilitating the lexicographer's job even more. In any case, the subsequent inversion will proceed much smoother if as many adequate L1 equivalents as possible are assigned to each sense of the respective L2 lemmata.

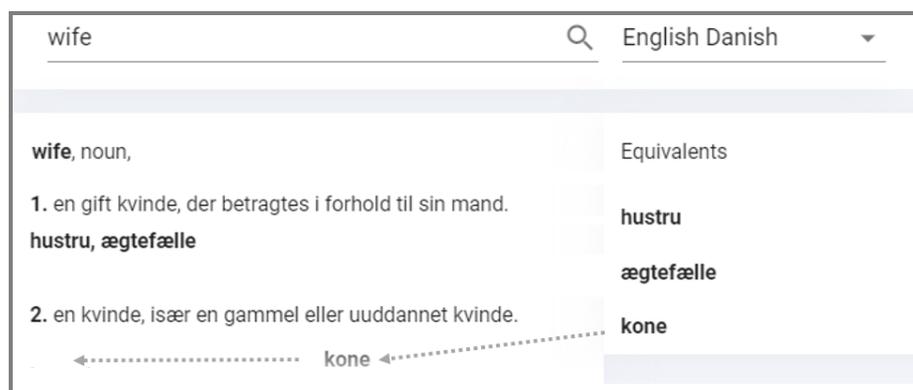


Figure 7: Assigning suggested L1 equivalents to senses

In this second test, we also learned more about the requirements for the original L2 definitions. Most of them worked perfectly, but others, like the first definition of *dragon*, were more problematic:

a mythical monster like a giant reptile. In European tradition the dragon is typically fire-breathing and tends to symbolize chaos or evil, whereas in the Far East it is usually a beneficent symbol of fertility, associated with water and the heavens.

This text is perfect as a lexicographical definition of *dragon*. But it is simply too wordy if it also has to function as a meaning differentiator in an L1–L2 diction-

ary, which is one of the possible products where the lexicographical data can be used. Pedro Fuertes-Olivera had similar challenges in his bilingual Valladolid project. The first small text segment (*a mythical monster like a giant reptile*) would serve very well as a differentiator. In this specific case, the remaining text provides additional information contrasting Eastern and Western culture. Other wordy definitions also have a similar structure with a small initial text segment followed by a longer one with more detailed information. In these cases, the problem seems to be the storage of the two parts, which does not allow them to be uploaded separately. The solution is here a case for prediction and interdisciplinary planning of databases. However, some wordy definitions do not have a similar internal structure and, thus, need to be modified accordingly. The Valladolid experience shows that it is possible to meet this requirement after some lexicographical training.

5. Some challenges

The tests show that machine translation can play a relevant role in the compilation of lexicographical databases and that the method presented in Section 3.4 to inverse the relationship between lemmata and equivalents is implementable. It suggests that overall the project is a success. It confirms the importance of never separating the L2 words from their definitions. However, the test of a relatively small amount of data does not provide answers to all questions. Many challenges remain, some of them predictable, others still to be detected. Here, I will briefly comment on two of the predictable ones.

One of these challenges is related to the stock of L1 lemmata once the inversion has taken place. For the benefit of future users, the most frequently consulted L1 lemmata must be part of this stock. The easiest way to verify if this is the case is to use log files from previous consultations. Although some L1 lemmata may be missing, the corresponding L2 senses will probably appear in the database in most cases. This shows the extraordinary importance of assigning as many L1 equivalents as possible to each sense of the L2 lemmata, preferably when preparing the L2–L1 part and alternatively when revising the L1–L2 part. Even so, some relevant single-word or multi-word L1 lemmata or their senses may still be absent. It is premature to develop the precise methods to solve these challenges based on the relatively limited experience from the above tests. It will have to wait until much more empirical data is available, probably when inverting a whole L2–L1 part. In any case, this is one of the tasks where the skilled lexicographer still has an important role to play.

Another predictable challenge is to determine to what extent existing data can be reused in a new dictionary project. There is no general answer to this question. Each publishing house will have to decide for itself. It will depend on the quality of the lexicographical data that is already available in its databases.

6. The human lexicographer's new role

As I wrote in Section 1, one of the objectives of my project at Ordbogen was to know more about the changing relationship between the human and what I call the artificial lexicographer. The latter refers to digital software that completes either compilation tasks, which the human lexicographer previously carried out, or entirely new tasks. It is not surprising at all that the human lexicographer's role in the compilation process changes. Something similar has happened throughout the history of lexicography whenever disruptive technologies have been applied. However, the current situation differs from past experiences in that the artificial lexicographer is increasingly replacing the human lexicographer and redefining his or her role. The main tendency today is that the human lexicographer dedicates more time to revising lexicographical data and less time to creative activities. This tendency seems inescapable. The project described above is no exception. Here, I will briefly list the tasks that are either new or different from what is common in other projects:

Core L2 database

- This is the part that involves the most creative work. Apart from the technological improvements that characterize all monolingual projects, the only thing new is the requirement for wordy definitions. These items should be structured in two parts to be stored separately in the database.

Bilingualized L2–L1 part

- checking translated definitions and reworking those that are erroneous or inadequate.
- checking translated example sentences and reworking those that are erroneous or inadequate.
- assigning as many L1 equivalents as possible to each sense of the L2 lemmata, either from a list of suggested candidates or using other methods, including introspection.

Inverted L1–L2 part

- checking the inversion in general, lemma by lemma.
- checking whether the most frequently consulted L1 lemmata appear after the inversion and whether their most relevant senses have L2 equivalents assigned.
- adding frequently consulted L1 lemmata when they are missing.
- assigning L2 equivalents to relevant L1 senses without equivalents.
- possibly checking translated example sentences to verify that they include the pertinent L1 lemma and not another L1 lemma (see Figure 6).

For the moment, this is all. But there may be other tasks to perform, or the ones mentioned may be slightly modified, when more experience is available.

7. Postscript

After testing the idea developed in Section 3.4, both Henrik Hoffmann and I believe it has a future. Machine translation has proven useful in lexicography, and inversion has become possible. But we are aware that there is still some work to do, both on the technological front and the lexicographical front. The experience shows that interdisciplinary collaboration is the best way to achieve fast and robust results. As stated in the introduction, I feel privileged to have had the opportunity to integrate the R&D team at Ordbogen. I urge other lexicographers to engage in similar forms of collaboration. The discipline is in great need of innovation in many aspects. The potential of the new technologies must be thoroughly explored. New ideas and hypotheses must be tested without fear of temporary setbacks. Everything useful must be incorporated to improve the compilation, storage, and presentation of lexicographical data. Nobody can do this alone.

Acknowledgements

I wish to thank the School of Communication and Culture, Aarhus University, for facilitating my research stay, and Ordbogen A/S for accepting my project and incorporating me into its R&D team.

Special thanks are due to Programmer and Web Developer Henrik Hoffmann for his patience, constructive comments, and technical assistance that allowed me to test and further develop my ideas.

Special thanks are also due to Prof. Rufus H. Gouws, Stellenbosch University, and Prof. Pedro Fuertes-Olivera, Valladolid University, for their willingness to discuss and comment on my research project, both during and after the stay at Ordbogen.

Finally, I also wish to thank Prof. Heming Yong, Guangdong University of Finance, Prof. Patrick Leroyer, and Prof. Martin Nielsen, both Aarhus University, for revising sample translations into Chinese, French, and German, respectively.

References

A. Machine translators

DeepL Translator: <https://www.deepl.com/translator>.

Google Translate: <https://translate.google.co.za/>.

B. Other literature

- Adamska-Sałaciak, A.** 2010. Why We Need Bilingual Learners' Dictionaries. Kernerman, I.J. and P. Bogaards (Eds.). 2010. *English Learners' Dictionaries at the DSNA 2009*: 121-137. Tel Aviv: K Dictionaries.
- Adamska-Sałaciak, A. and I. Kernerman.** 2016. Introduction: Towards Better Dictionaries for Learners. *International Journal of Lexicography* 29(3): 271-278.
- Bergenholtz, H., I. Cantell, R.V. Fjeld, D. Gundersen, J.H. Jónsson and B. Svensén.** 1997. *Nordisk Leksikografisk Ordbok*. Oslo: Universitetsforlaget.
- Bothma, T.J.D.** 2011. Filtering and Adapting Data and Information in an Online Environment in Response to User Needs. Fuertes-Olivera, P.A. and H. Bergenholtz (Eds.). 2011. *e-Lexicography: The Internet, Digital Initiatives and Lexicography*: 71-102. London/New York: Continuum.
- Burkhanov, I.** 1998. *Lexicography. A Dictionary of Basic Terminology*. Rzeszów: Wydawnictwo Wyższej Szkoły Pedagogicznej.
- Duda, W.** 1986. Ein "aktives" russisch-deutsches Wörterbuch für deutsch-sprachige Benutzer? Günther, E. (Ed.). 1986. *Beiträge zur Lexikographie slawischer Sprachen*: 9-15. Berlin: Akademie-verlag.
- Fuertes-Olivera, P.A. and S. Tarp.** 2014. *Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminography*. Berlin/Boston: De Gruyter.
- Fuertes-Olivera, P.A. and S. Tarp.** 2020. A Window to the Future: Proposal for a Lexicography-assisted Writing Assistant. *Lexicographica* 36: 257-286.
- Fuertes-Olivera, P.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.
- Gellerstam, M.** 1999. LEXIN — lexikon för invandrare. *LexicoNordica* 6: 3-18.
- Hartmann, R.R.K. and G. James.** 1998. *Dictionary of Lexicography*. London/New York: Routledge.
- Hausmann, F.J.** 1977. *Einführung in die Benutzung der neufranzösischen Wörterbücher*. Tübingen: Niemeyer.
- Lew, R. and A. Adamska-Sałaciak.** 2015. A Case for Bilingual Learners' Dictionaries. *ELT Journal* 69(1): 47-57.
- Marello, C.** 2003. The Bilingual Dictionary: Definition, History, Bidirectionality. Hartmann, R.R.K. (Ed.). 2003. *Lexicography. Critical Concepts II*: 325-342. London/New York: Routledge.
- Martínez de Sousa, J.** 1995. *Diccionario de Lexicografía Práctica*. Barcelona: Bibliograf.
- Mikkelsen, H.K.** 1992. What Did Scerba Actually Mean by "Active" and "Passive" Dictionaries? Hyldgaard-Jensen, K. and A. Zettersten (Eds.). 1992. *Symposium on Lexicography V. Proceedings of the Fifth International Symposium on Lexicography May 3-5, 1990, at the University of Copenhagen*: 25-40. Tübingen: De Gruyter Mouton.
- Scerba, L.V.** 1940 (1995). Towards a General Theory of Lexicography. *International Journal of Lexicography* 8(4): 314-350.
- Tarp, S.** 2005. The Concept of a Bilingual Dictionary. Barz, I., H. Bergenholtz and J. Korhonen (Eds.). 2005. *Schreiben, Verstehen, Übersetzen, Lernen. Zu ein- und zweisprachigen Wörterbüchern mit Deutsch*: 27-41. Frankfurt am Main: Peter Lang.
- Tarp, S.** 2008. *Lexicography in the Borderland between Knowledge and non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's lexicography*. Tübingen: Max Niemeyer.

- Tarp, S.** 2013. New Developments in Learner's Dictionaries III: Bilingual Learner's Dictionaries. Gouws, R.H., U. Heid, W. Schweickard and H.E. Wiegand (Eds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*: 425-430. Berlin: De Gruyter.
- Tarp, S.** 2019. Connecting the Dots: Tradition and Disruption in Lexicography. *Lexikos* 29: 224-249.
- Tarp, S., K. Fisker and P. Sepstrup.** 2017. L2 Writing Assistants and Context-Aware Dictionaries: New Challenges to Lexicography. *Lexikos* 27: 494-521.
- Tomaszczyk, J.** 1983. On Bilingual Dictionaries: The Case for Bilingual Dictionaries for Foreign Language Learners. Hartmann, R.R.K. (Ed.). 1983. *Lexicography: Principles and Practice*: 41-51. London: Academic Press.