Interactive Word Sense Disambiguation in Foreign Language Learning

Jasper Degraeuwe
LT^3 / MULTIPLES
Ghent University
Belgium
Jasper.Degraeuwe@UGent.be

Patrick Goethals
LT^3 / MULTIPLES
Ghent University
Belgium
Patrick.Goethals@UGent.be

Abstract

"Word sense awareness" is a feature which is not yet implemented in most corpus query tools, Intelligent Computer-Assisted Language Learning (ICALL) environments or computer-readable didactic resources such as graded word lists (Alfter and Graën, 2019; Pilán et al., 2016; Tack et al., 2018). The present paper aims to contribute to filling this lacuna by presenting a word sense disambiguation (WSD) method for ICALL purposes. The method, which is targeted at Spanish as a foreign language (SFL), takes a few prototypical example sentences as input, converts these sentences into "sense vectors", and integrates part of the training data collection process into interactive vocabulary exercises. The evaluation of the method is based on a selection of 50 ambiguous items related to the domain of economics and compares different types of input data. With a top weighted F1 score of 0.8836, the present study shows that the currently available NLP tools, resources and methods provide all the necessary building blocks for developing a WSD method which can be integrated into interactive ICALL environments.

1 Introduction

Compared to single-meaning words, lexically ambiguous items (e.g. *empleo*: 'usage' / 'job') have shown to be more challenging to process and learn (Bensoussan and Laufer, 1984; Degani and Tokowicz, 2010). Nevertheless, the distinction of word senses has often been overlooked in the design of vocabulary learning curricula and graded word lists (Tack et al., 2018). Moreover, when foreign language teachers or textbook designers need a set of usage examples for each sense of an ambiguous word, they often have to manually gather or invent these example sentences. Or, if they are able to use corpus query tools, they have to rely on concordance searches which do not distinguish between word senses, as most of those tools only allow performing searches on word forms.

It is for these kinds of time-consuming tasks that the field of ICALL aims to offer solutions: by means of Natural Language Processing (NLP)-driven methodologies, ICALL studies seek to facilitate and/or (partially) automate the creation of language learning materials to be used in a CALL environment. To tackle the lexical ambiguity issue, the NLP technique of WSD can be applied (Kulkarni et al., 2008). Although performance levels have recently breached the "80% glass ceiling set by the inter-annotator agreement" (Bevilacqua et al., 2021), WSD is still an open problem (Blevins et al., 2021; Navigli, 2018), especially for languages other than English and for specific purposes such as ICALL. However, thanks to the recent advances within NLP, the tools and resources to successfully develop an ICALL-tailored WSD method do seem to be available. Therefore, with this study we aim to make a plea for integrating WSD in ICALL, presenting a straightforward method which can easily be implemented in existing ICALL environments.

The paper is structured as follows: Section 2 first of all zeroes in on the concepts of lexical ambiguity (as conceived in NLP) and WSD, and also provides a brief overview of the recent developments within ICALL. Next, in Section 3 we present our WSD method, which is aimed at Spanish as the target language (3.1), takes a few prototypical example sentences as input (3.2), leverages the ability of Transformer models to create contextualised "sense vectors" (3.3), and integrates part of the process of compiling training data into interactive vocabulary exercises for SFL students (3.4). The WSD method is applied to and evaluated on custom datasets (3.5), the results of which are discussed in Section 4. Finally, Section 5 includes a conclusion and discussion of the study, alongside some possible directions for future research.

Jasper Degraeuwe and Patrick Goethals. Interactive word sense disambiguation in foreign language learning.
Proceedings of the 11th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2022). Linköping Electronic Conference Proceedings 190: 46–54.

2 Related research

2.1 Lexical ambiguity in NLP

In the domain of (written) NLP, a lexically ambiguous item is usually defined as a lemma of a specific part of speech (POS) for which more than one sense can be distinguished. For reasons of feasibility and scalability, to determine which senses are included in the sense inventory (i.e. the lexicon in which ambiguous words are linked to their different senses), most computationally-focused studies on WSD rely on established resources such as (Euro)WordNet (Fellbaum, 1998) or BabelNet (Navigli and Ponzetto, 2012). However, the sense distinctions in these resources are often of a very fine-grained nature, which makes them sometimes even difficult for humans to distinguish (Loureiro et al., 2021) and, in many cases, unsuitable for real-life NLP applications (Hovy et al., 2013). Moreover, Kilgarriff (1997) argues that "there is no reason to expect a single set of word senses to be appropriate for different NLP applications", since "different corpora, and different purposes, will lead to different senses".

In other words, our specific ICALL setting requires a specific sense inventory, tailored to the needs of SFL learners (see Section 3.2). An example of an inventory with coarse-grained sense distinctions that are easily interpretable by humans is the CoarseWSD-20 dataset (Loureiro et al., 2021), which consists of a manual expert selection of twenty English nouns and their corresponding senses, and is based on Wikipedia as reference inventory and corpus. Degraeuwe et al. (2021) undertake a similar effort, but in this case to build a WSD system which distinguishes between sensory and non-sensory meanings of ambiguous items for the specific purpose of analysing the use of sensory language as a rhetoric technique in tourism discourse.

2.2 Word sense disambiguation

As formulated by Navigli (2009), WSD is "the ability to computationally determine which sense of a word is activated by its use in a particular context". Formally, this means that WSD aims to identify a mapping A from words to senses (i.e. to assign the appropriate sense(s) to all or some of the words in a text), such that $A(i) \subseteq SensesD(w_i)$, where $SensesD(w_i)$ is the set of senses encoded in a dictionary D (i.e. the sense inventory) for word w_i, and $A(i)$ is that subset of the senses (usually of length 1) of w_i which are appropriate in the context (Navigli, 2009). In the following example, a WSD system is expected to map *operación* in sentence (a) to the sense "operation", and in sentence (b) to the sense "surgery".

(a) La **operación** supuso la transferencia de cerca de 500 trabajadores. ('The operation entailed transferring around 500 workers.')

(b) La **operación** se ha efectuado por medio de un cateterismo. ('The surgery has been performed by means of a catheterisation.')

WSD can be conceived as a classification task, with the word senses as the classes, and an automatic classification method as the means to assign each occurrence of a word to one or more classes based on the evidence from the context and/or from external knowledge sources. In this regard, it should be highlighted that, contrary to other NLP classification tasks such as POS tagging and Named Entity Recognition (NER), in WSD there is no fixed number of predefined categories (classes), since the set of senses (classes) is different for each individual word. In other words, "WSD actually comprises n distinct classification tasks, where n is the size of the lexicon" (Navigli, 2009). As a result, building a WSD system usually constitutes an accumulative process.

2.3 WSD in ICALL

Driven by the recent advances in NLP, current ICALL applications which can be used for vocabulary learning purposes are doing more and more credit to the "Intelligent" part of their name. In the category of intelligent corpus consultation applications, the hybrid HitEx system for Swedish (Pilán et al., 2016) is a well-known example: it allows extracting context-independent example sentences of a given proficiency level from corpora by performing fine-grained and customisable queries. To this end, the system relies on computer-readable lexical-semantic resources and POS-tagged, lemmatised and parsed Swedish corpora, to which then a series of rule-based and machine learning-based selection criteria are applied. Next, for the category of exercise generation applications, different examples are to be found in the work of Graën, whose research explores the use of (multi)parallel corpora as input data for the automatic generation of (gamified) language learning

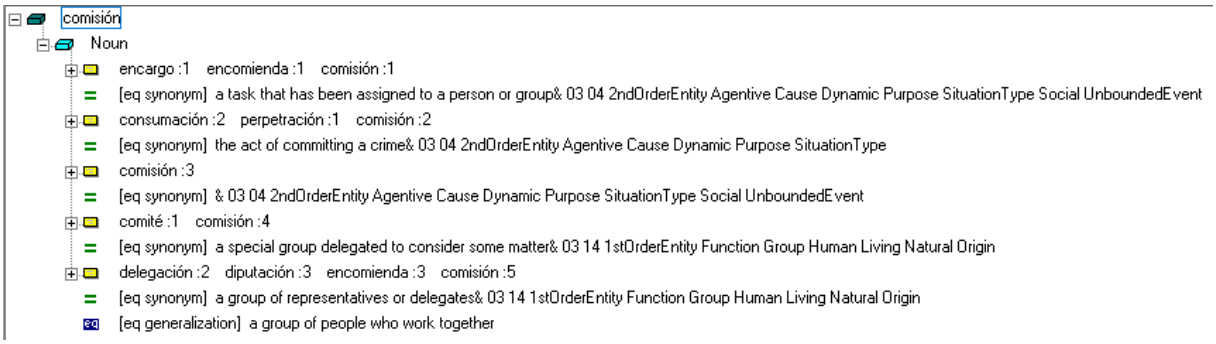

Figure 1: Spanish WordNet entry for *comisión*, in which two of its five "synsets" (synonym sets) refer to the sense "committee" with hardly any difference between them: the synset *[comité, comisión]* and the synset *[delegación, diputación, encomienda, comisión]*. Furthermore, despite being used very frequently, *comisión* as "intermediary fee" is not included amongst the senses.

exercises, ranging from training knowledge of particle verbs (Alfter and Graën, 2019) to reordering exercises (Zanetti et al., 2021).

However, although this kind of systems have proven to be a valuable complement to vocabulary learning activities in the classroom (Ruiz et al., 2021), using ICALL still comes with its limitations. Recognising lexically ambiguous items and distinguishing between their senses is one of those pending issues (Pilán et al., 2016), as the NLP-driven technique of WSD is rarely integrated in ICALL environments, in corpus query tools or in the development of computer-readable resources for didactic purposes (e.g. graded word lists).

3 Methodology

3.1 Setting

As mentioned in the introduction, one of the novel aspects of our WSD method is its embedding in an educational context. For this study, we take a B2+ level Spanish writing course at university as the target setting. As a part of the vocabulary learning module of the course, which specifically focuses on learning business vocabulary, the 35 enrolled students work with the ICALL environment of the Spanish Corpus Annotation Project (SCAP; scap.ugent.be; Goethals, 2018) and have to complete an online module on lexical ambiguity. It is in that module on lexical ambiguity that part of the training process of our WSD system is integrated (Section 3.4).

To arrive at a selection of target items the WSD method can be applied to and tested upon, all nouns in a 11M corpus containing newspaper articles on economics, are first ranked from highest to lowest keyness compared to a reference corpus (both corpora are available within the SCAP platform), with the keyness calculation being performed according to the Log Ratio formula (Hardie, 2014). Next, we ask an SFL expert to select the first 50 items (see Table 2) which have at least two relatively frequent meanings and fit within the business vocabulary scope of the B2+ writing course.

3.2 Sense inventory

Since using existing resources such as the Spanish WordNet and BabelNet would result in a too-fine grained and sometimes incomplete inventory (see Figure 1 for an example), we elaborate a custom sense inventory based on the senses included in the Spanish dictionary Clave. [1] Given its status as a general dictionary and its focus on "contemporarily used expressions and terms in daily life" (Fundación SM, 2021), Clave provides suitable input for building an SFL-focused sense inventory. To build the actual contents of the inventory, we ask an SFL expert to go over the Clave senses and, if deemed necessary, group related senses together into coarse-grained "main senses". In addition, the expert is instructed to eliminate all domain-specific Clave senses which are not related to the domain of economics (e.g. *matriz* as "matrix" in the domain of mathematics). Importantly, for most of its senses, the Clave dictionary provides a prototypical usage example, which will be used as the input data of our WSD methodology. If no example sentence is available for a given main sense (which is the case for 16.5% of the main senses), a usage example taken from one of the SCAP cor-

[1] Complete sense inventory available at https://github.com/JasperD-UGent/sense-inventory-economics-50.

Unseen sentence to be classified		
Eso sí, tendrás que aprobar también el examen de **ingreso**. ('Of course, you will have to pass the entrance exam.')		
Senses	**Labelled example sentences**	**Cosine similarity**
Sense 1 "entry"	Apoyaremos tu **ingreso** en la comisión. ('We will support your entry into the commission.')	.5591
	Hoy a las seis de la tarde es el **ingreso** del nuevo académico. ('Today at six in the afternoon the inauguration of the new academic takes place.')	**.5626**
Sense 2 "deposit"	El **ingreso** puedo realizarlo en cualquier sucursal. ('I can make the deposit in any branch.')	**.5026**
Sense 3 "income, revenue"	Este mes, los **ingresos** han sido menores porque ha habido menos ventas. ('This month, revenue has been lower because there have been fewer sales.')	**.3893**

Table 1: Authentic application example of the cosine similarity classifier, with the maintained cosine similarity values put in bold. The predicted output for the unseen sentence containing the ambiguous item *ingreso* is "entry", as the highest maintained value corresponds to this sense.

pora is manually added.

3.3 Sense vectors

Next, for each of the 50 target items, the prototypical example sentences included in the sense inventory are transformed into "sense vectors". To this end, we take the contextualised word embedding of the ambiguous item in the sentences with the help of the RoBERTa-BNE model (Gutiérrez-Fandiño et al., 2021). As a result, each main sense in the sense inventory is now represented by a set of n unique vectors, where n is the number of prototypical sentences linked to the main sense (see Figure 2 for an example). Usually, n is equal to 1, but if multiple Clave senses have been grouped together n can also be greater than 1. Finally, the sense vectors are used to predict the correct sense of ambiguous instances in new, unseen sentences. To perform this classification task, we use cosine similarity calculations, a measure closely related to distance metrics such as the Euclidean distance (which is used in k-NN classifiers), with the main difference being that instead of the distance between two vectors, it is the cosine of the angle between them which is measured. Cosine similarity calculations usually yield outcome values between 0 (no similarity) and 1 (complete similarity), and can be used to rank relative similarity levels (i.e. higher scores indicate a higher level of similarity).

In summary, given a new target sentence with an ambiguous word, the individual cosine similarity values between the vector of the ambiguous item in this target sentence and its sense vector(s) are computed. Next, only the highest cosine similarity value for each sense is maintained, after which the classifier assigns the target sentence to the sense with the highest maintained value (see Table 1 for an example).

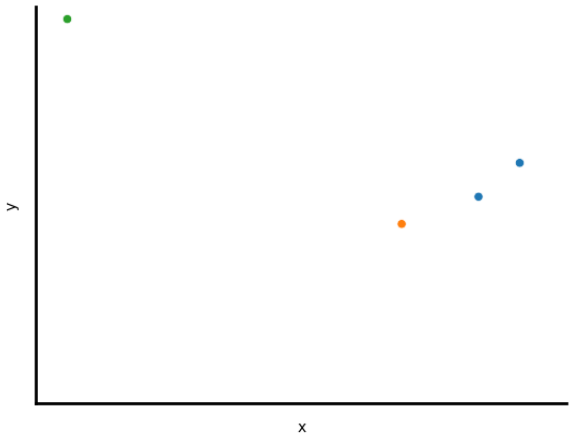

Figure 2: Authentic example of sense vectors visualised in a two-dimensional space, for the item *ingreso* (see Table 1 for the sentences used to create the vectors). The blue dots correspond to sense vectors of the sense "entry", the orange dot to "deposit", and the green one to "income, revenue".

3.4 Interactive exercises for training

As mentioned in Section 3.1, we use an online module on lexical ambiguity included in an SFL writing course at university to compile additional training data. To this end, for each of the 50 selected target items, a series of interactive exercises are elaborated in which the 35 SFL students enrolled in the course familiarise themselves with the linguistic phenomenon of lexical ambiguity and

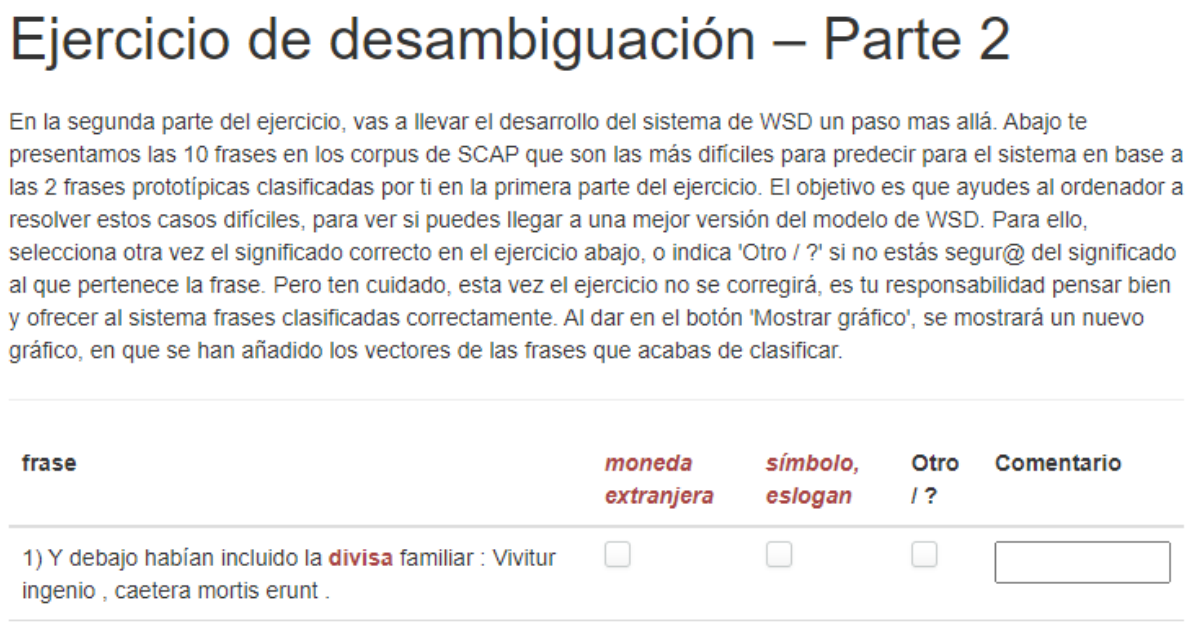

Figure 3: Screenshot of the all-embracing exercise in which students can train their own WSD model, for the item *divisa* ('foreign currency' / 'symbol, motto'). Before arriving at this part of the exercise, students first had to initialise their WSD model by assigning the prototypical example sentences from the sense inventory to the right sense. These labelled sentences were then converted into sense vectors and used to identify the ten most difficult sentences for the system (i.e. the ten sentences with the lowest cosine similarity difference between the two top maintained values) in a selection of unseen sentences taken from the SCAP corpora. In the exercise part shown in the screenshot, students are asked to assign these ten sentences to the correct sense, in order to provide the system with additional training data. Once finished, students are brought to the final part of the exercise, in which they can analyse the performance of their custom model on new sentences.

learn the different meanings of the ambiguous vocabulary item in question. Towards the end of the exercise series, students are also encouraged to consider lexical ambiguity from the perspective of the computer, and receive a brief introduction into the NLP technique of WSD. Finally, as an all-embracing exercise, they are offered the opportunity to train their own WSD models. Amongst other activities, this final exercise consists of assigning 10 unseen ambiguous sentences of a given target item to the correct sense (see Figure 3). These exercise responses are collected in order to be used as additional training data.

As all students are pre-assigned 8 vocabulary items for which they have to complete the entire exercise series (with the vocabulary items being evenly distributed across the students), for every vocabulary item at least 5 responses can be collected for each of the 10 unseen ambiguous sentences. Finally, a threshold-based filter is applied to the gathered data: all sentences for which at least 80% of the responses have been assigned to the same sense are considered suitable to be used as additional training data for that particular sense.

3.5 Evaluation

Since our WSD method is designed to be applied in a foreign language learning setting, it could not be evaluated using one of the (few) existing WSD datasets for Spanish (e.g. Màrquez et al., 2004). First of all, many of the 50 vocabulary items selected from the economic target corpus do not occur amongst the ambiguous words included in these datasets. Working with the words of the existing datasets instead of selecting the target items ourselves would have solved this problem, but none of the datasets includes a set of ambiguous items which could serve as input for the real-life vocabulary class as described in Section 3.1. Moreover, most datasets are labelled according to WordNet sense distinctions, which were not designed for the purpose of foreign language learning. In other words, all annotations would first have had to be manually converted to the sense distinctions made in our SFL-tailored inventory before they would become usable.

Therefore, we decide to create custom datasets, based on data from the SCAP corpora. For each of the 50 selected ambiguous items, all sentences

in which the lemma of the item occurrs are extracted from the corpora. For this concordance search query, the minimal sentence length was put to 10 and the maximal length to 70, to ensure that noisy data are being kept out (e.g. short phrases with a lack of contextual information and paragraphs in which sentence splitting was not performed correctly). The resulting datasets per ambiguous item are then cleaned following an automatic, rule-based process, and randomly split into a 100-sentence test set and a "rest set" with all remaining sentences. Finally, the test sets are manually annotated by an SFL expert according to the sense distinctions made in the custom sense inventory.

To evaluate both the WSD method in general and the added value of using exercise responses as additional training data, two different input types (i.e. the training data which are used by the WSD system to make predictions) are determined. To make the system more robust, the first step of both input types consists of automatically identifying, for each sense of each ambiguous item, the 10 instances in the "rest set" with the highest cosine similarity compared to the contextualised sense vectors included in the sense inventory (see Section 3.3). The vectors corresponding to the ambiguous item in those sentences are then added as extra labelled training data on top of the original sense vectors. In the basic input type ("base"), no other training data are added after this step. In the second input type ("enriched"), however, the selected sentences from the interactive vocabulary exercise (see Section 3.4) are included as additional training instances.

Finally, the WSD method is applied twice to the test sets, once for every input type. To measure performance, weighted F1 scores are calculated: this score represents the harmonic mean of precision (i.e. the number of truly positive predictions divided by the number of truly positive and falsely positive predictions) and recall (i.e. the number of truly positive predictions divided by the number of truly positive and falsely negative predictions). By using the weighted variant of the metric, unequal label distributions are balanced out.

4 Results

First of all, the average results presented in Table 2 show that both input types outperform the most frequent sense (MFS) baseline by a large margin, highlighting the overall potential of the WSD method. Since, to the best of our knowledge, no benchmark exists for WSD for language learning purposes, to interpret the F1 scores we compare our results to Loureiro et al. (2021), a study with a similar setup as ours (see also Section 2.1). On a dataset of 20 English nouns, the fine-tuned large BERT model of Loureiro et al. (2021) obtains a top weighted F1 score of 0.975. However, it should be highlighted that they make use of labelled training sets with sizes up to 6421 instances. In this regard, the scores achieved by the best-performing model in our methodology, which only takes a few sentences as labelled input, can be considered highly satisfactory. Next, the results also reveal that the addition of the exercise responses as additional training data ("enriched") leads to a 0.01 increase in performance. Clearly, this increase is too small to make firm claims about the added value of resolving the most difficult cases (recall that the examples to be classified by the students correspond to the examples with the lowest cosine similarity difference between the two top maintained values) and adding them as training data.

As for the individual results, the scores reveal a mixed picture. First, for some items (*asociación*, *cuota*, *déficit*, *emisión*, *explotación* and *operación*) the addition of the exercise responses appears to cause a reverse effect. Although these non-neglegible decreases in performance are balanced out by the considerable improvements for *balance*, *comisión*, *compañía*, *descuento*, *división*, *entidad*, *gestión*, *ingreso*, *matriz*, *participación* and *valoración*, this finding suggests that new example sentences should be added with caution. When checking the added sentences, for *asociación*, *cuota* and *explotación* we found one or two sentences to be classified incorrectly by the students, which could explain part of the lower F1 score for those words. For the other items, resolving the most difficult cases seems to introduce "confusion" rather than clarity into the system. This finding could be an indication that we might need to reconsider the choice for taking this type of examples as our source for new training data. Switching to the exact opposite starting point, for instance, could be another approach worth studying: instead of integrating the sentences with the smallest cosine similarity differences into vocabulary exercises, the sentences with the largest differ-

Individual results							
Ambiguous item (Log Ratio)	**#senses**	**F1_base**	**F1_enriched**	**Ambiguous item (Log Ratio)**	**#senses**	**F1_base**	**F1_enriched**
acción (4.5)	4	.9909	.9547	entidad (8.5)	2	.8904	.9411
administración (7.1)	3	.909	.8623	explotación (6.1)	2	.9604	.8937
aplicación (5.4)	4	.8635	.8621	facturación (10.2)	3	.7997	.8372
área (5.1)	2	.8435	.8435	firma (5.7)	3	.8417	.8838
asociación (5.5)	2	.962	.9083	gestión (7)	2	.8877	.9809
balance (6.3)	2	.7755	.8493	implantación (6.4)	3	.8513	.8899
bono (9.3)	2	.9156	.9454	ingreso (6.8)	3	.9016	.9697
colocación (4.5)	3	.9417	.93	inversión (9)	3	.7226	.7664
comisión (6.8)	4	.9295	.9833	liquidación (6.7)	3	.7709	.775
compañía (5.5)	4	.7709	.9054	matriz (6.1)	3	.8547	.9622
competencia (6)	2	.9591	.949	mercado (7.2)	2	.9463	.964
concesión (5.1)	3	.9402	.9402	operación (6.2)	2	.9483	.8913
cotización (8.4)	3	.8129	.849	operador (8.7)	3	.7539	.7191
crecimiento (8.6)	2	.7786	.8211	organismo (5.1)	2	.9535	.9619
cuota (6.6)	2	.8719	.7231	participación (6.5)	2	.7863	.864
déficit (6.7)	2	.9804	.863	plataforma (4.5)	5	.9491	.9491
demanda (6.5)	2	.8698	.897	política (5.2)	2	.8368	.8368
descuento (6.8)	2	.9111	.9655	préstamo (5.7)	2	.9198	.9198
deuda (5)	2	.7048	.7379	rebaja (5.5)	2	.9482	.9482
distribución (6.4)	2	.9168	.8949	saneamiento (5.3)	2	.8024	.8024
divisa (5.6)	2	.9663	.9569	sector (6.8)	3	.8912	.8912
división (5.4)	6	.7146	.8376	segmento (8.8)	2	.9295	.9295
ejercicio (5.3)	4	.9259	.9172	subida (5)	2	.9879	.9879
emisión (7.3)	4	.8693	.7133	tasa (8.1)	3	.9218	.9218
empleo (5)	2	1	1	valoración (5.5)	2	.4713	.5806
Average results							
F1_base	.873						
F1_enriched	**.8836**						
MFS	.5901						

Table 2: Performance results on the custom 100-sentence test sets. The individual results report the weighted F1 scores for each item with "base" and "enriched" as the two different input types. Log Ratio values are added between brackets. For the average results, the mean of all 50 individual scores is taken. Here, also the most frequent sense (MFS) baseline is reported, a simple but often hard-to-beat dummy system which always predicts the most frequent sense of the ambiguous item (which was identified as the most frequent sense amongst the test set annotations).

ences could be taken as input for a new type of exercise. Finally, the individual results also highlight that a few items appear to be particularly challenging for the system (e.g. *valoración*: 'estimate' / 'appreciation, evaluation'), and will need to receive special attention. In this regard, a possible addition to the methodology could be to calculate the cosine similarity between the original sense vectors in order to determine an "inter-sense similarity" score. If, for a given ambiguous item, this score exceeds a certain threshold, the item could then be flagged so that more example sentences can be added before initialising the WSD method.

5 Conclusion and discussion

In this study, a novel WSD methodology for ICALL purposes is presented, applied to Spanish as the target language. The method makes use of a customised sense inventory in which all senses are accompanied by one or a few prototypical example sentences. By means of the RoBERTa-BNE model (Gutiérrez-Fandiño et al., 2021), these sentences are converted into unique "sense vectors", which can then be introduced into the cosine similarity classifier to predict the sense of an unseen ambiguous instance. Finally, we study the embedding of part of the training process into interactive vocabulary learning exercises for SFL students.

To assess performance, the method is applied to custom datasets for a selection of 50 ambiguous nouns related to the domain of economics. Overall, the WSD system achieves very promising results, with a top average weighted F1 score of 0.8836. Next, compiling additional training data through interactive vocabulary exercises leads to a 0.01 increase in performance compared to not using the exercise responses as additional training data. As the increase is only of a very small nature, additional research with a larger number of target items and/or larger test sets will be required to reach well-founded conclusions on this particular aspect. Finally, the analysis of the individual performance results indicates that adding the exercise responses does not per se lead to improved performance, especially (and perhaps logically) when incorrect classifications by the students passed the 80% threshold. Nevertheless, as more and more exercise responses will be collected over time, more sentences can be added (which could mitigate the "confusion" that is sometimes introduced) and more responses per sentence can be gathered (which could enable us to apply a more strict threshold for selecting suitable sentences). Additionally, switching to another type of input sentences in the exercises (e.g. the least difficult sentences instead of the most difficult ones) could also be a path worth exploring.

As the language model used to create the vectors is pretrained (and can thus be used off the shelf) and the exercise responses are filtered in an automated fashion, the prototypical example sentences are the only manually curated data needed to initialise the methodology. This architecture makes the WSD method scalable and applicable in real-life scenarios. Therefore, with this research we hope to contribute to implementing the distinction of word senses as an additional feature in corpus query tools, ICALL environments or computer-readable resources for didactic purposes (e.g. graded word lists), which would open a wide range of opportunities for the design of different language learning materials. These materials can range from lexical-semantic resources in which ambiguous items with similar polysemy patterns are grouped together, over disambiguated graded vocabulary lists, to exercises which start by presenting the so-called core meaning of polysemous items, a type of exercise which has proven to be beneficial for the long-term retention of those items (Verspoor and Lowie, 2003).

However, future research will still need to address the detection of low-performing items, and study how the performance of these items can be improved. For example, the cosine similarity between the original sense vectors could be calculated to determine an "inter-sense similarity" score. If, for a given ambiguous item, this score exceeds a certain threshold, the item could then be flagged. Similarly, the agreement rates between students on the interactive exercises can also be taken as a measure to detect possibly challenging items: if exercise responses show little consensus this should perhaps not be considered as a lack of inter-annotator agreement, but rather as a sign that (some of) the sense distinctions of the ambiguous word might be particularly challenging. Thirdly, we plan to carry out a follow-up study with a larger number of target items and multiple SFL students as test set annotators, and make the corresponding datasets publicly available so that they can be used to benchmark WSD methods for ICALL purposes. Finally, we also aim to expand our coverage to verbs and adjectives, which will likely entail other challenges given their different syntactic and morphological characteristics.

Acknowledgments

This research has been carried out as part of a PhD fellowship on the IVESS project (file number 11D3921N), funded by the Research Foundation - Flanders (FWO). For the elaboration of the custom sense inventory, we relied on the Clave dictionary, to whose contents we have access thanks to a research collaboration with the Fundación Santa María.

References

David Alfter and Johannes Graën. 2019. Interconnecting lexical resources and word alignment: How do learners get on with particle verbs? In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 321–326.

Marsha Bensoussan and Batia Laufer. 1984. Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, 7:15–32.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent Trends in Word Sense Disambiguation: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 4330–4338,

Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

Terra Blevins, Mandar Joshi, and Luke Zettlemoyer. 2021. FEWS: Large-Scale, Low-Shot Word Sense Disambiguation with the Dictionary. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 455–465, Online. Association for Computational Linguistics.

Tamar Degani and Natasha Tokowicz. 2010. Ambiguous words are harder to learn. *Bilingualism: Language and Cognition*, 13(3):299–314.

Jasper Degraeuwe, Patrick Goethals, and Pauline Verhoeve. 2021. Ampliar la caja de herramientas del análisis del discurso asistido por el ordenador: el caso de los cinco sentidos en el discurso turístico. *Les Cahiers du GERES*, 12:91–109.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Fundación SM. 2021. Diccionario Clave. Lengua española. https://www.grupo-sm.com/es/book/diccionario-clave-lengua-espa%C3%B1ola.

Goethals, Patrick. 2018. Customizing vocabulary learning for advanced learners of Spanish. In *Technological innovation for specialized linguistic domains : languages for digital lives and cultures, proceedings of TISLID'18*, pages 229–240. Éditions Universitaires Européennes. Event-place: Gent, Belgium.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2021. MarIA: Spanish Language Models. Publisher: arXiv Version Number: 5.

Andrew Hardie. 2014. Log Ratio: An informal introduction. *ESRC Centre for Corpus Approaches to Social Science (CASS)*, pages 1–2.

Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Adam Kilgarriff. 1997. I don't believe in word senses. *Language Resources and Evaluation*, 31(2):91–113.

Anagha Kulkarni, Michael Heilman, Maxine Eskenazi, and Jamie Callan. 2008. Word Sense Disambiguation for Vocabulary Learning. In Beverley P. Woolf, Esma Aïmeur, Roger Nkambou, and Susanne Lajoie, editors, *Intelligent Tutoring Systems*, volume 5091, pages 500–509. Springer Berlin Heidelberg, Berlin, Heidelberg. ISSN: 0302-9743, 1611-3349 Series Title: Lecture Notes in Computer Science.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and Evaluation of Language Models for Word Sense Disambiguation. *Computational Linguistics*, pages 1–57.

Lluis Màrquez, Mariona Taulé, Antonia Martí, Núria Artigas, Mar García, Francis Real, and Dani Ferrés. 2004. Senseval-3: The Spanish lexical sample task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 21–24, Barcelona, Spain. Association for Computational Linguistics.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Roberto Navigli. 2018. Natural Language Understanding: Instructions for (Present and Future) Use. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 5697–5702, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Ildikó Pilán, Elena Volodina, and Lars Borin. 2016. Candidate sentence selection for language learning exercises: From a comprehensive framework to an empirical evaluation. *Revue Traitement Automatique Des Langues*, 57(3):67–91.

Simón Ruiz, Patrick Rebuschat, and Detmar Meurers. 2021. The effects of working memory and declarative memory on instructed second language vocabulary learning: Insights from intelligent CALL. *Language Teaching Research*, 25(4):510–539.

Anaïs Tack, Thomas François, Piet Desmet, and Cédrick Fairon. 2018. NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 137–146, New Orleans, Louisiana. Association for Computational Linguistics.

Marjolijn Verspoor and Wander Lowie. 2003. Making Sense of Polysemous Words. *Language Learning*, 53(3):547–586.

Arianna Zanetti, Elena Volodina, and Johannes Graën. 2021. Automatic Generation of Exercises for Second Language Learning from Parallel Corpus Data. *International Journal of TESOL Studies*, 3:55–70.