

XD-AT: A Cross-Dictionary Annotation Tool

Meritxell González, Charlotte Buxton, Roser Saurí

Oxford University Press, United Kingdom

Abstract

Linking lexical datasets to each other is a key strategy for expanding and enriching their content with additional data from other resources. However, different resources show significant differences in the degree of granularity of the lexicographic information. Thus, while extending more coarse-grained datasets with content from fine-grained ones seems a feasible task, the other way around cannot be tackled directly. For this reason, linking datasets at the level of meaning rather than word level is essential. But also, for the same reason, word alignment at the level of meaning is a challenging task not yet solved. Within this context, we created XD-AT, a web-based annotation tool aimed to assist humans to annotate linked sense pairs across dictionaries. In this work, we focus in XD-AT's main functionalities, capabilities and potential extensions, such as reusability and adaptability. For example, although XD-AT has been implemented to classify the type of relationship between linked senses from an English monolingual dictionary and the English side of bilingual ones, XD-AT can also be extended into a more general annotation tool for marking up any type of cross-dictionary mappings at the sense level.

Keywords: annotation tool; sense linking; dictionary mark-up; meaning overlap.

1 Introduction

This work presents XD-AT, a web-based annotation tool for marking up relations across dictionaries taking place at the sense level. XD-AT has been developed within the framework of Prêt-à-Llod, an EU funded project devoted to developing multilingual linked data and language technology (i.e., Linguistic Linked Open Data). The context of XD-AT within the project is particularly concerned with the exploration of methodologies for linking dictionaries at the level of meaning. This is been an area of significant activity in the past decade with the successful linking of key lexical databases, such as WordNet or Wikipedia, for language technology purposes. Examples of such success are BabelNet (Navigli & Ponzetto, 2012) and the work summarized in Gurevych, Eckle-Kohler & Matuschek (2016). Efforts in this area have more recently turned into the alignment of dictionary content due to the benefits that dictionary sense linking can contribute. Very significantly, it opens up the possibility of expanding existing lexicographic content with additional data from other sources, for example for building specialized multilingual lexicons (Schmidt, 2009), or for creating new bilingual dictionaries (Gracia et al., 2019; McCrae, et al., 2017; Saurí et al. 2019).

Currently there are a number of dictionary writing systems (DWSs) that are used to generate and edit dictionary content (cf. Abel, 2012). Some are quite well-known off-the-shelf proprietary solutions, such as T-LEX,¹ and IDM DPS², while others have been developed within the open source paradigm, like DEBWrite (Rambousek & Horák, 2015) and Lexonomy (Měchura, 2017). Similarly, there already exist text annotation tools (TATs), some of which are specifically for sense tagging text, such as WebAnno (Eckart de Castilho et al., 2016), STSAnno (Batanovićet al. 2018), and Ubyline (Miller et al. 2016). However, to the best of our knowledge, there is no tool among either DWSs or TATs that provides the functionality for cross-dictionary sense alignment annotations. For example, for mapping senses of different dictionaries that refer to the same meaning, or for qualifying whether aligned senses differ in any way.

XD-AT aims to supply such functionality, which involves addressing very specific challenges. Similar to DWSs, our tool had to be able to display in a clear, differentiated way, the several parts in the structure of a dictionary entry, such as definitions, example sentences, and labels. Moreover, similar to TATs, the system had to facilitate the classification of the targeted annotation elements given a closed set of classes (e.g., sense-link vs. non-sense-link). In addition, the system had to extend beyond the functionality of both DWSs and TATs in order to be able to allow for annotations on pairs of (as opposed to single) dictionary units, which entails a degree of structural (and therefore layout) complexity due to the hierarchical nature of dictionary information. Therefore, XD-AT addresses this gap among annotation tools for language resources of different kinds.

To tackle XD-AT's development, we focus on a very specific use case: to mark-up differences in sense granularity between two linked senses from different dictionaries in order to, with the resulting annotations, train a machine learning-based classifier able to determine these distinctions automatically. The annotation strategy defined by lexicographers for that task guided XD-AT functional requirements. For example, we wanted it to allow for multiple annotators on the same data in order to avoid any annotator bias and also to be able to compute inter-annotation agreement (IAA). Other requirements that were essential for us were: easy access, clear information layout, and user-friendliness. All these were taken into account in the design and deployment of the tool, together with the goal of enabling its reusability in other cross-dictionary sense annotation tasks in the future.

¹ <https://tshwanedje.com/tshwanelex/>

² <https://www.idmgroup.com/content-management/dps-info.html>

2 Sense alignment and granularity differences

This section presents the specificities of our particular annotation use case with the aim of facilitating the understanding later on, of the requirements that drove the design and deployment of the tool. As said, the goal was to obtain manual annotations on granularity differences between the two dictionary senses in a sense link. We define *sense link* as a pair of senses, each from a different dictionary, which represent the same meaning for a given *lexeme*. For lexeme we understand a combination of a lemma and a lexical category. For example, *water* (noun) and *water* (verb) are two different lexemes. By contrast, what in a dictionary may be considered as independent lexical items (e.g., homographs like *lie* (verb) ‘to not tell the truth’ and *lie* (verb) ‘to adopt or be in a horizontal position’) will be taken here as belonging to the same lexeme. When aligning senses for a lexeme in one dictionary with the equivalent senses for the same lexeme in another dictionary, we can see that in some cases they fully align (i.e., they refer exactly to the same meaning), whereas in others one of the senses (or both) extends beyond the meaning conveyed by the other. This is illustrated in Figure 1 below, which presents the senses for lexeme *fog* (noun) from an English monolingual dictionary (left), and its translation into Spanish from a bilingual dictionary (right). As can be observed, sense 1.2 in the monolingual perfectly aligns with sense 2 in the bilingual, while sense 1 in the bilingual covers the meaning of both senses 1 and 1.1 in the monolingual.

EN Monolingual	EN-ES
fog NOUN 1. [mass noun] A thick cloud of tiny water droplets suspended in the atmosphere (...) 1.1 [in singular] An opaque mass of particles in the air. 1.2 <i>Photography</i> Cloudiness which obscures the image on a developed negative (...) 2. [in singular] A state or cause of perplexity or confusion.	fog noun UNCOUNTABLE AND COUNTABLE 1. (<i>Meteorology</i>) niebla (feminine) 2. (<i>Photography</i>) velo (masculine)

Figure 1: Senses for lexeme *fog* (noun) in a monolingual (left) and bilingual (right) dictionary.

Taking into account these differences in the semantic extent and overlap of two senses, we defined the four different types of meaning relationship between two linked senses illustrated in Table 1: *perfect*, *narrower-than*, *wider-than*, and *partial*.³ *Perfect* match indicates that each sense aligns completely throughout the full extension of the other one. In other words, the entire meaning expressed by one sense is also expressed and covered by the other one. In contrast, *narrower-than* and *wider-than* account for sense pairs where a sense in one dictionary has a broader meaning than the sense in the other dictionary. This occurs when the meaning of one sense fully overlaps with the other one but does not fully enclose it (*narrower-than*), or the other way around (*wider-than*). Finally, *partial* is used when each of the two senses’ meaning extends beyond the reference of the other and thus, there is a part of the meaning covered by each sense that is not included in the other one.

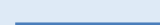



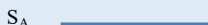

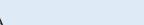

	Perfect match	Different sense granularity		Different sense boundaries
Meaning alignment	S_A  S_B 	S_A  S_B 	S_A  S_B 	S_A  S_B 
Grounding relationships	S_A fully overlaps with S_B and encloses it.	S_A fully overlaps with S_B but does not enclose it.	S_A partially overlaps with S_B and encloses it.	S_A partially overlaps with S_B and does not enclose it.
Sense link classes	Perfect	Narrower-than	Wider-than	Partial
Symbol	=	<	>	~

Table 1: Types of sense link based on differences in sense granularity.

In order to conduct the human classification task with confidence, annotators therefore needed access to the information related to both dictionary senses as well as the possibility of selecting among the four values just presented. In addition, we wanted to meet the specifications for first-class annotation tools (e.g., user friendliness, support for managing the annotation tasks, ability to deal with multiple annotations, etc.). In the following sections, we explain the requirements we identified and how they were deployed into XD-AT.

3 General Framework Specifications

Web-based access. XD-AT is a web-based application with database support. One of the main advantages of web applications is that they allow working from any device connected to the Internet using a browser, without the need to install additional software. Although this was not a hard requirement for our experiment, we consider it would bring a valuable benefit for future re-usability and scalability of the tool, especially in the particular set-up where external contributors may participate in the annotation task.

³ This classification has also been adopted in McCrae, ELEXIS Monolingual Word Sense Alignment Task (2020)

User roles. XD-AT is also a multi-user application that requires authentication. The current version defines three types of user roles: *annotators*, *judges* and *managers*, and a synthetic role named *automatic*. Only managers can create new users (either other managers or annotators), handle the assignment of annotation tasks across users, inspect the annotation progress, and enable the judge review of certain assignments. In turn, annotators can only see their own assignments and progress. The judge is the user role created to resolve classification discrepancies when there are multiple annotations for the same sense link; and the automatic role is used to store classifications produced by automatic means (e.g. by an algorithm).

Data storage. The entire information handled by XD-AT is stored in a relational database. This covers the user credentials mentioned above, the dictionary information to be displayed, the organization of sense links collections into batches, batch assignments to annotators, all annotations and re-annotations produced by annotators, judges and automatic roles, as well as application-specific definitions, such as the list of lexical categories, polysemy degrees, and sense link classes. The latter is indeed a relevant feature in XD-AT. On the one side, it simplifies the adoption of the tool by others who can reuse the pre-defined closed set of possible annotation labels, and annotation tasks characteristics. On the other side, it still allows decoupling the application-specific characteristics from the tool implementation, which endows the application with higher adaptability to other annotation tasks that may use a different set of classes or prefer to organize the annotation tasks using different criteria.

Next sections give more details of all these functionalities, and how they are displayed in the interface.

4 Annotation Panel

The information to be displayed on the interface was carefully selected. It was important to be able to provide annotators with all the information needed to ground their decisions, but also that the display was as light as possible to avoid visual stress during the annotation work.

Dictionary information. For correctly classifying a sense link, annotators required access to all the other existing links for either sense in the targeted link. Figure 2 shows a screenshot of the annotation panel. The left side displays information for the monolingual dictionary, while the right one shows the piece for the bilingual counterpart. The area in between the two frames displays symbols for the different types of links the annotator must choose from: = (perfect match), < (narrower-than), > (wider-than), ~ (partial match), ? (donotknow), unlink (for cases that in fact are not links). Additional sense links for the senses in the targeted link are shown in two further frames at the lower part of the frame, along with their current type class (in green in between the panels).

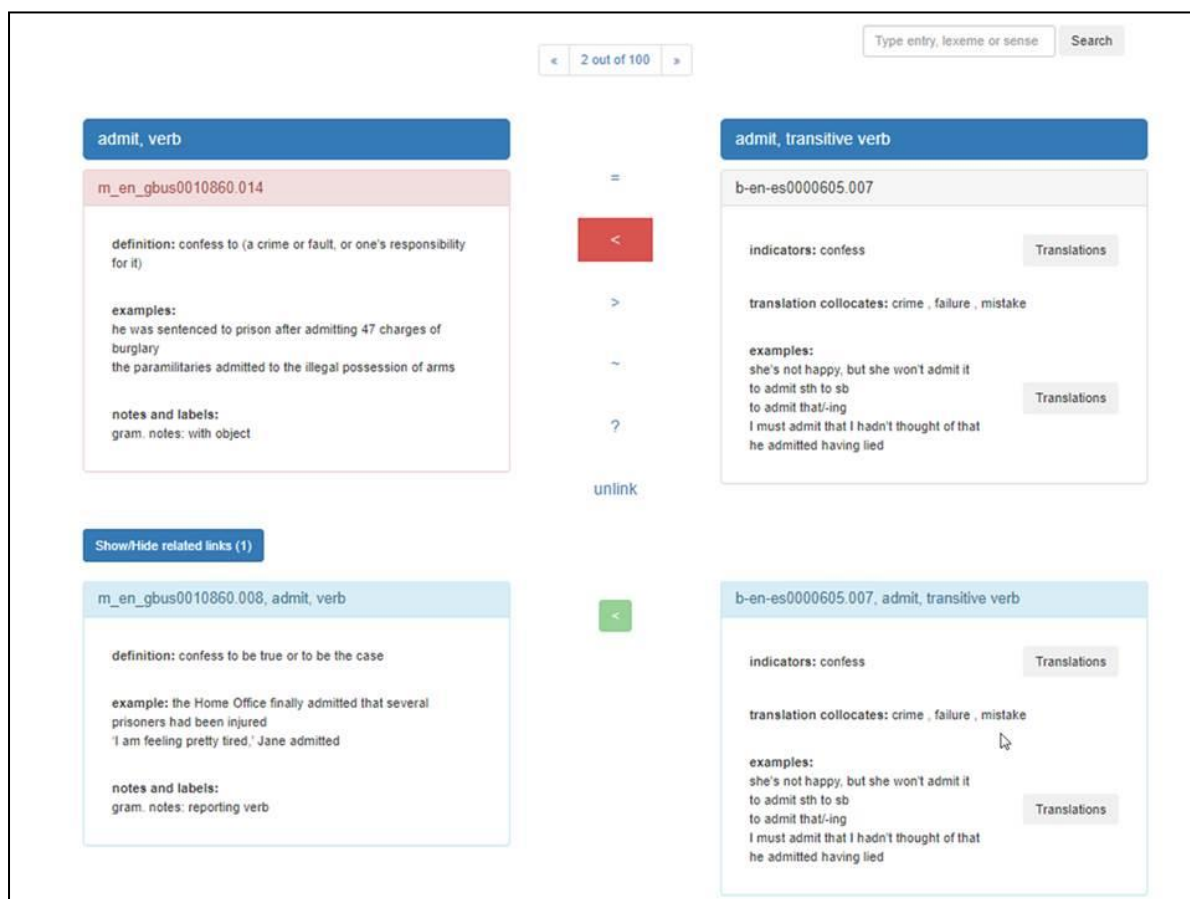


Figure 2: Screenshot of the annotation panel.

For the monolingual dictionary, the interface shows definitions, examples, grammatical⁴ and technical notes,⁵ word forms,⁶ domain, region and register labels,⁷ and domain and semantic classes from a taxonomy.⁸ Similarly, for the bilingual dictionary, it shows indicators,⁹ collocates,¹⁰ examples, and the additional labels and notes described above. Translations were also included but hidden because annotators pointed out that they were not needed in most of the cases. Thus, the interface included a button that shows lemma's and examples' translations on demand. As a matter of fact, results from the experiments in Kouvara et al. (2020) supported the hypothesis that translations are not essential for this particular task (although they may help to discern some difficult cases). In our experiments involving bilingual dictionaries for Spanish, Chinese and Russian, none of the annotators knew either Russian nor Chinese, but some knew Spanish. However, the inter-annotators agreement shows little differences among the three datasets.

Search functionality. Annotators did not consider it essential to have access to the whole sense inventory for the linked lexeme in either dictionary, and therefore that information was not included as part of the default display in order to avoid visual clutter. Instead, the tool includes a search engine that allows retrieving this information for any given substring or identifier, in case annotators are interested in looking it up. The goal is to assist them by providing them access to any piece of information they may need without having to reach to the dictionary sources externally from XD-AT. Such a functionality is only possible because of having stored the dictionary data in a relational database.

Pre-annotated labels. To facilitate annotators work, the system was designed so that it was able to offer annotators with a pre-annotated choice (the one estimated as the most likely) for them to validate. More specifically in the context of our project, sense links were already pre-classified based on the set of heuristics (Kouvara et al., 2020), which in turn were based on the number of links held by each sense in either side of the link (i.e. in either dictionary). Hence, the annotator task consisted in correcting, or confirming, the class pre-assigned to each link. To help annotators to distinguish automatic labels from those that had already been corrected, we created the colour scheme shown in Figures 3a and 3b: red was used for automatically computed labels (via the role *automatic*), and green for manually annotated ones (given by the annotator). That way, it was visually easy to identify on which links the annotator had already taken a decision.

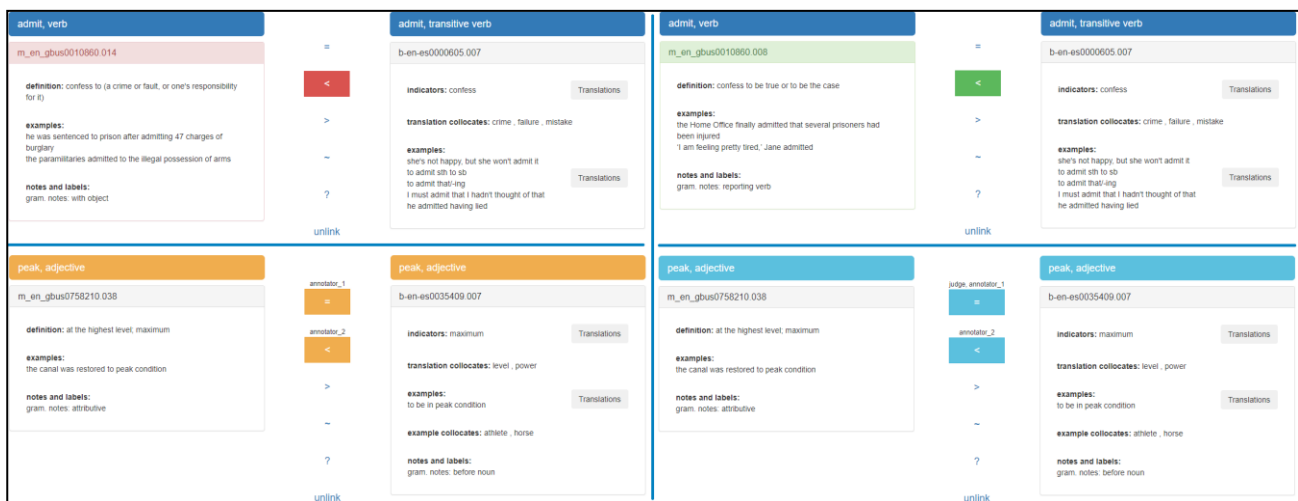


Figure 3: XD-AT colour scheme: a) automatic class (red), b) human label (green), c) disagreement between annotators (orange), d) reviewed by the judge (blue).

Judge disagreements review. XD-AT also includes a judge review mechanism to resolve disagreements between annotators and assign a final label in those cases. Given a particular batch, it finds all sense links for which there is at least one classification difference among the annotators, and shows them all in an annotation panel designed for that purpose (Figures 3c and 3d). The color scheme in this view consists of orange for displaying all annotators' choices, and light blue for indicating the judge's final choice. Note also the small text on top of the annotation label buttons. These are the annotator names along with their choice. This view cannot be seen by annotators so as to avoid them influencing each other's decisions.

⁴ In Figure 2, grammatical notes for both senses of *admit* (verb) in the monolingual: with object and reporting verb.

⁵ Technical notes for *sodium hydroxide* (noun): 'Chemical formula: NaOH'.

⁶ Word forms for *world* (noun): [usually] 'the world'.

⁷ Register labels for *think big* (idiomatic): [informal]; region label for *barbie* (noun): [chiefly Australian, New Zealand]; domain label for *arteriosclerosis* (noun): [Medicine].

⁸ Semantic class for *arteriosclerosis* (noun): [physiological state]; and domain class: [Pathology].

⁹ In Figure 2, *admit* (verb) has a single indicator: 'confess'.

¹⁰ In Figure 2, *admit* (verb) has three collocates: 'crime, failure, mistake'.

5 Annotation Task Management

XD-AT organizes annotation data into batches, which are sets of annotation units (e.g., here, sense links) of a specific length. The notion of batch has been a great enabler for better organizing and managing the annotation task. In our case, batches consisted of 100 sense links each.

Batch creation. An important feature in XD-AT is the automatic creation of batches of data to annotate based on carefully selected criteria. Batches are sets of annotation units (here, sense links). For our case in particular, the full collection of links was split into lexeme subsets according to their lexical category (noun, verb, adjective, adverb/preposition, or other) and polysemy degree (single-sense, small, medium, or large size). The rationale behind that is that senses for certain lexical categories and, most likely, polysemy degrees may be harder to annotate than others. Thus, the aim with the batch creation functionality is to help annotators focus on similar annotation cases at a time.

Batch assignment. Once batches are created, a manager user needs to assign them to annotators. Ideally, batches could be evenly distributed and automatically assigned among annotators, so that each annotator is assigned a similar number of batches of each type. However, we kept this as a manual operation so that the manager can adapt assignments to the annotators skills to increase the quality of the results.

Finally, XD-AT features two further functionalities to manage and monitor the annotation task progress: assignment of multiple annotations and annotation history track. They are presented next.

Multiple annotation assignment. Batches can be assigned to several annotators at once in order to obtain multiple annotations on the same data. This is a key feature because it supports several functions: firstly, it makes it possible to calculate IAA as an estimate on the difficulty of the task; secondly, it allows us to identify areas of major disagreement among annotators (and therefore presumed data complexity) that can then inform the development of well-grounded annotation guidelines; and last but not least, it helps pinpoint annotators that tend to disagree with others more often, a piece of information that can then be used to adjust batch assignment in order to ensure highest data accuracy.

Annotation history track. Finally, XD-AT stores all re-annotations over each sense link; that is, all the labels that the same annotator may have assigned to a sense link at different moments in time due to second thoughts or hesitation about that case. The re-annotation history is useful to analyze data complexity and task difficulty. Among other things, it can help identify common lexicographic features among annotation units that create more difficulties, or pinpoint particular batches that are more challenging than others.

6 Exporting Annotated Data

XD-AT is able to export the annotations in a machine-readable format (CSV and JSON) according to three use cases:

- **Baseline:** These are the labels created by automatic roles. In our downstream experiments, we use these to compare the accuracy of different machine learning models trained on the manual annotations, hence the name.
- **Shared annotations:** This is the collection of labels assigned by all annotators to sense links in the shared batches (i.e., the batches adjudicated to multiple annotators). This subset is used for, e.g., computing IAA or identifying areas of major disagreement. It does not include judge labels.
- **Gold standard:** This includes the final classification labels for all the batches in the dataset. In the case of shared batches, it takes the judge's decisions in case of disagreement among annotators. This subset discards all links classified as unlink or uncertain.

In our specific annotation task, the export output contains the following information fields: 1) sense link type, which is the final label only; 2) polysemy degree of the lexeme that the linked senses belong to; 3) lexical category of that same lexeme; 4) dictionary to which each of the linked senses belongs; 5) annotation batch ID; 6) annotation timestamp; and 7) annotator ID. The goal is to provide as much information as possible along with annotations, so that it can be used in downstream tasks.

Finally, XD-AT also exports the list of links that were re-annotated, grouped by annotator and batch ID. This information, together with the list of links labelled as donotknow, can be used to further analyse the complexity of the task, discern patterns in difficult cases, and also guide enhancements of XD-AT in future revisions.

7 Final Remarks

XD-AT was developed for annotating distinctions of sense granularity between dictionary senses that refer to the same meaning (that is, that are already aligned). However, it can be upscaled into a more general tool for also marking up any type of cross-dictionary alignments and relations at the sense level.

We are not aware of any other tool developed so far for that purpose. Possible extensions include:

- Improving user management functionalities, e.g. use of other authentication methods, ability to enable/disable users, assignment of multiple roles to the same user, etc.
- Facilitating the analysis of IAA scores online. For example, by adding interface areas for inspecting the inter-annotator agreement results in an interactive way, by selecting the batches (or users) among which to compare annotations.
- Improving export functionalities by, e.g., including interface areas for selecting batches or dictionaries subsets to export, or for sampling based on users or other fields, in addition to the classification labels.
- Publishing the tool publicly. To date, XD-AT is for internal use only since the above extensions are work in progress.

Nonetheless, we are open to receive requests for using the tool and suggestions for making XD-AT a more flexible tool able to embrace other use cases.

8 References

- Abel, A. (2012). Dictionary Writing Systems and Beyond. A S. Granger, & M. Paquot (Ed.), *Electronic Lexicography* (p. 83–106.). Oxford, United Kingdom: Oxford University Press.
- Kouvara, E., González, M., Grosse, J., Sauri, R. (2020). Determining Differences of Granularity between Cross-Dictionary Linked Senses. *Congress of the European Association for Lexicography (Euralex 2020)*. Alexandroupolis, Greece.
- Batanović, V., Cvetanović, M., & Nikolić, B. (2018). Fine-grained Semantic Textual Similarity for Serbian. *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)* (p. 1370-1378). Miyazaki, Japan: ELRA.
- Eckart de Castilho, R., Mújdricza-Maydt, É., Muhie Yimam, S., Hartmann, S., Gurevych, I., Frank, A., & Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)* (p. 76-84). Osaka, Japan: The COLING 2016 Organizing Committee.
- Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378-382.
- Fleiss, J., Levin, B., & Cho Paik, M. (2003). *Statistical Methods for Rates and Proportions* (3rd ed.). Hoboken, New Jersey: Wiley.
- Gracia, J., Kabashi, B., Kernerman, I., Lanau-Coronas, M., & Lonke, D. (2019). Results of the Translation Inference Across Dictionaries 2019 Shared Task. *Proceedings of TIAD-2019 Shared Task*. 2493, p. 1-12. Leipzig, Germany: CEUR Workshop Proceedings.
- Gurevych, I., Ecker-Kohler, J., & Matuschek, M. (2016). *Linked Lexical Knowledge Bases: Foundations and Applications*. Morgan & Claypool.
- McCrae, J. (May / 2020). *ELEXIS Monolingual Word Sense Alignment Task*. CodaLab Competition: <https://competitions.codalab.org/competitions/22163>
- McCrae, J., Bond, F., Buitelaar, P., Cimiano, P., Declerck, T., Gracia, J., Piasecki, M. (2017). 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets. co-located with 1st Conference on Language, Data and Knowledge (LDK 2017). *Proceedings of the LDK 2017 Workshops*. 1899. Galway, Ireland: CEUR Workshop Proceedings.
- Měchura, M. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. A I. Kosem, C. Tiberius, M. Jakubiček, J. Kallas, S. Krek, & V. Baisa (Ed.), *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*, (p. 662-679). Leiden: Lexical Computing.
- Miller, T., Khemakhem, M., Eckart de Castilho, R., & Gurevych, I. (2016). Sense-annotating a Lexical Substitution Data Set with Ubyline. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (p. 828–835). Portorož, Slovenia: European Language Resources Association (ELRA).
- Navigli, R., & Ponzetto, S. (December / 2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, 217-250.
- Rambousek, A., & Horák, A. (2015). DEBWrite: Free Customizable Web-based Dictionary Writing System. *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 Conference*. (p. 443-451). Herstmonceux Castle: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.
- Sauri, R., Mahon, L., Russo, I., & Bitinis, M. (2019). Cross-Dictionary Linking at Sense Level. *42nd Conference on Very Important Topics (CVIT 2016)*, 15.
- Schmidt, T. (2009). The Kicktionary – A Multilingual Lexical Resource of Football Language. A H. Boas (Ed.), *Multilingual FrameNets in computational lexicography : methods and applications* (p. 101-132). Berlin, Germany: de Gruyter.

Acknowledgements

This work has been funded by the H2020 project “Prêt-à-LLOD: Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors” under grant agreement No 825182. Also, we are very grateful to Ashleigh Alderslade, Matthew Bladen, Emma Davies, Janet Gough, Denny Hilton, Eleanor Maier, Iona Ogilvie, Nick Rolfe, and Catherine Sangster, the expert lexicographers who contributed all the editorial knowledge we were lacking, and also helped to test the first version of the tool and annotate the data for our experiments. In addition, we would like to thank Eva Theodoridou and Anna Emberton for their support in managing and planning the work required for the project. The authors are responsible for any errors and problems.