M A T E M A T I C K O - F Y Z I K Á L N Í F A K U L T A

P R A H A

Annotation guidelines for Spanish verbal synonyms in the SynSemClass Lexicon

Cristina Fernández-Alcaina, Eva Fučíková, Zdeňka Urešová

ÚFAL Technical Report
TR-2022-72

ISSN 1214-5521

U N I V E R S I T A S C A R O L I N A P R A G E N S I S

Copies of ÚFAL Technical Reports can be ordered from:

Institute of Formal and Applied Linguistics (ÚFAL MFF UK)
Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, CZ-11800 Prague 1
Czechia

or can be obtained via the Web: https://ufal.mff.cuni.cz/techrep

Annotation guidelines for Spanish verbal synonyms in the *SynSemClass* Lexicon

Cristina Fernández-Alcaina, Eva Fučíková, Zdeňka Urešová
alcaina|fucikova|uresova@ufal.mff.cuni.cz

Acknowledgements

The work described herein has been supported by the Grant Agency of the Czech Republic under the EXPRO programme as project "LUSyD" (project No. GX20-16819X) and uses resources hosted by the LINDAT/CLARIAH-CZ Research Infrastructure (project No. LM2018101, supported by the Ministry of Education of the Czech Republic). The German part of the lexicon is partly supported by the grant Humane AI Network, funded by the EC by award No. 952026.

Abstract

These guidelines are created as part of the research project *Language Understanding: from Syntax to Discourse* (LUSyD) and describe the goals, resources and procedure for the addition of a fourth language, Spanish, to the multilingual synonym verbal lexicon *SynSemClass*. The lexicon currently comprises three languages: Czech, English and German, thus allowing a comparison of syntactic-semantic cross-linguistic properties. The lexicon also provides empirical data relevant for ongoing NLP research.

Contents

PART 1: ANNOTATION

1.1 Introduction

The *SynSemClass* lexicon[1] classifies contextually-based synonymous verbs into multilingual synonym classes based on the semantic and syntactic properties they display. For each class, the lexicon provides a set of semantic roles (Roleset) mapped to the valency arguments identified for each class (Role_Argument mapping). The name of the class in each language corresponds to the verb considered to represent the prototypical sense. Each individual verb included in a given class, i.e., Class Member (henceforth, CM) is linked to a group of existing online lexical resources available for each language, e.g., Vallex for Czech, VerbNet for English, E-VALBU for German, AnCora for Spanish, among others. Each CM is exemplified by instances of real texts (and their translations) extracted from parallel corpora. Specifically, data is extracted from the *Prague Czech-English Dependency Treebank* (PCEDT)[2] for Czech-English, from the *Paracrawl*[3] corpus for German-English and from the *XSRL* dataset (Daza & Frank 2021) for Spanish-English.

What follows is a description of the annotation of Spanish verbs in the *SynSemClass* lexicon. These guidelines are partly based on the method followed for the annotation of Czech, English and German in previous versions of the lexicon[4].

1.2 Goal of the annotation

The aim of this project is to expand the multilingual lexicon *SynSemClass* by adding Spanish synonymous verbs. As in previous versions of the lexicon, data have been collected from a sentence-aligned parallel corpus. For Spanish, verbs have been extracted from the *X-SRL* dataset[5] (Daza & Frank 2021). The corpus consists of original texts from the English portion

---

[1] The fourth version of the lexicon can be accessed here: https://lindat.mff.cuni.cz/services/SynSemClass/

[2] https://ufal.mff.cuni.cz/pcedt2.0/en/index.html

[3] https://opus.nlpl.eu/ParaCrawl.php

[4] For Czech-English, see Urešová, Zdeňka, Eva Fučíková & Eva Hajičová. *CzEngClass: Contextually-based Synonymy and Valency of Verbs in a Bilingual Setting.* Technical Report no. TR-2019-62, Copyright © ÚFAL MFF UK, Prague, Czechia, ISSN 1214-5521, 101 pp., 2019. Available at: https://ufal.mff.cuni.cz/techrep/tr62.pdf For German-English, see Urešová, Zdeňka, Eva Fučíková, Jan Hajič & Karolina Zaczynska. *Annotation guidelines for German verbal synonyms included in SynSemClass Lexicon*. Technical Report no. TR-2021-70, Copyright © ÚFAL MFF UK, Prague, Czechia, ISSN 1214-5521, 23 pp., 2021. Available at: https://ufal.mff.cuni.cz/techrep/tr70.pdf

[5] Daza, Ángel & Annete Frank. 2021. *X-SRL: Parallel Cross-lingual Semantic Role Labeling*. 1-58563-962-1. https://doi.org/10.35111/10zk-gq05 (Accessed via https://www.ldc.upenn.edu/).

of the 2009 *CoNLL Shared Task* Part 2 (Hajič et al. 2009)[6] automatically translated into Spanish using *DeepL*.

Synonymy is understood in terms of contextually-based synonymy. Thus a verb is considered to be a member of a class if it expresses the same meaning attested in the rest of the members of such class in a context, i.e., it has a similar valency to the verbs contained in the class. For example, for the class *pohltit*/*absorb*/*absorbieren* (vec00476), the arguments identified are 'ACT/A0' and 'PAT/A1'. If the requirement is fulfilled, the Spanish verb added to that class will be assigned the same set of semantic roles described for that particular class. For the example above, the roles assigned are 'Absorber' and 'Absorbed', respectively.

A specific annotation tool (*Synonyms Editor*, henceforth, *SynEd*) has been developed to facilitate the task of human annotators (Figure 1).

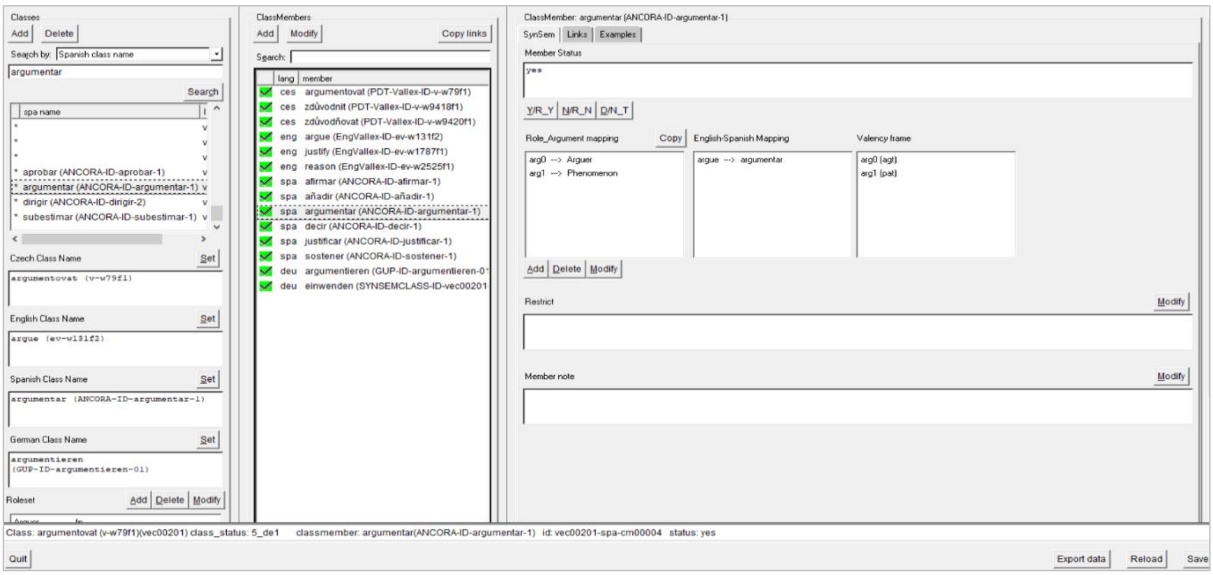

Figure 1. A screenshot of SynEd

The editor (from left to right) contains:

i) the verb classes (*Classes*) (including the name of each class in each language and the Roleset described for that class),

ii) the verbs contained in each class (*Class Members*),

iii) the semantic and syntactic properties of each class (*SynSem*),

iv) the links to other online lexical resources (*Links*), and

v) the examples extracted from corpora (*Examples*).

[6] Hajič, Jan, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, & Yi Zhang. 2009. The CoNLL-2009 shared task: syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task (CoNLL '09)*. Association for Computational Linguistics, USA, 1–18.

1.3 Annotation task

Since the annotation of Spanish verbs builds on existing annotations for Czech, English and German, both the multilingual verb synonym classes and the specific Roleset for each class are already defined. Each CM in a given class is described by:

i) a mapping of the semantic roles and the valency frame of the class,
ii) links to other online lexical resources, and
iii) a set of representative examples extracted from parallel corpora.

The task of the annotator is to go through the automatically generated Spanish verbal CMs candidates. For each CM, annotators are expected to:

i) Select the Class Member Status (Section 1.4)
ii) Map each of the roles provided in the Roleset to a specific argument (Role_Argument mapping, Section 1.5)
iii) Add restrictions or notes relevant for the annotation process (if applicable) (Section 1.6)
iv) If not added yet, set a verb as the Spanish class name (Section 1.7)
v) Add links to external online resources (Spanish FrameNet, AnCora, ADESSE, SenSem and Spanish WordNet) (Section 1.8)
vi) Select appropriate examples from the XSRL corpus (Section 1.9)

It is important to point out that the structure of these guidelines does not necessarily reflect the exact order of the steps followed in the annotation of synonymous verbs but rather how they are organized in the editor. Annotation is in fact a process with interlinked steps that usually requires to go backwards and forwards to decide whether a particular verb should be added to a class, the resources to which it should be linked and the set of examples that better captures the meaning.

1.4 Class Members status annotations

Class Member Status indicates whether a particular verb sense should be included or not in a class. In our view, meanings are determined by the valency frame displayed by a particular verb. Thus, it is possible to identify several senses for a single verb based on different valency frames, including antonym senses.

After inspection of the valency frame of the CM, the annotator must select one of the following options in the *Member Status* box of the *SynSem* tab:

i) Y/R_Y (*Yes*/*Rather_Yes*): the verb sense belongs or may belong to the class (i.e., it displays the same valency frame), or

ii) N/R_N (*No*/*Rather_No*): the verb sense does not belong or may not belong to the class (i.e., it has a different valency frame).

iii) D/N_T (*Delete*/*Not_Touched*): the verb is discarded from the class because it has been incorrectly assigned to the class (e.g., *pass* → *repercutir*), possibly due to a wrong pairing in a previous step in which verbs were automatically paired. [7]

To facilitate the identification of a verb as a member of a particular class, annotators can check the definition of each verb class by hovering over the verb which gives name to the class. The class name can be found on the left side of the editor (Figure 2).

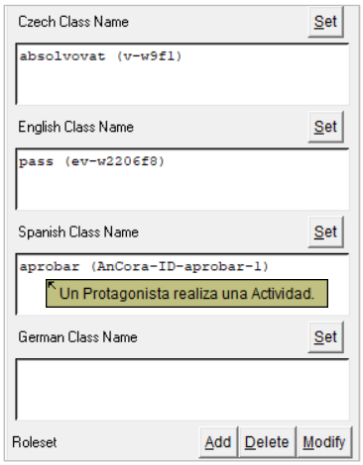

Figure 2. Definition (in Spanish) for the class *absolvovat/pass/aprobar* (vec00601)

1.5 Role_Argument mapping

As mentioned above, CMs are required to have the same or similar meaning and valency structure as the rest of the members already contained in the class. Therefore, for each CM of a certain synonymous class, each of the semantic roles in its Roleset needs to be meaningfully mapped to an element of its valency frame. Roles can be mapped to:

[7] Double click the corresponding option button to select the 'Rather_Yes', 'Rather_No' and 'Not_Touched' status.

i) either a member of the valency frame, e.g., arg0, arg1, argM, or

ii) a proxy functor: #sb (animate), #sth (inanimate), #any[8] (both animate and inanimate), #smh (somehow), #smt (sometime), #swh (somewhere).

Each obligatory member of the valency frame of a verb listed in the given synonymous class must be mapped to a semantic role from the assigned Roleset[9]. Annotators can check the definition of each role by hovering over the name of the role in the Roleset box[10], located on the left side of the editor (Figure 3).

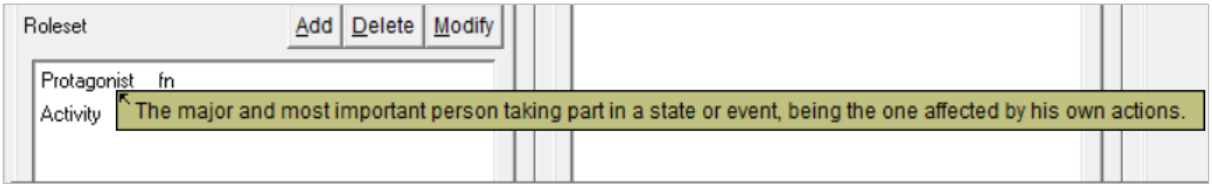

Figure 3. Role definitions in the Roleset of the class *absolvovat/pass/aprobar* (vec00601)

In SynEd, there are two types of verbs with two different IDs: i) AnCora-ID and ii) SynSemClass-ID. Sections 1.5.1 and 1.5.2 provide specific instructions on the role-argument mapping of each type.

1.5.1 *AnCora-ID*

The role-argument mapping is annotated under the Syntactic-Semantic Information tab in the editor (*SynSem*) and, specifically, in the *Role_Argument mapping* box (Figure 4).

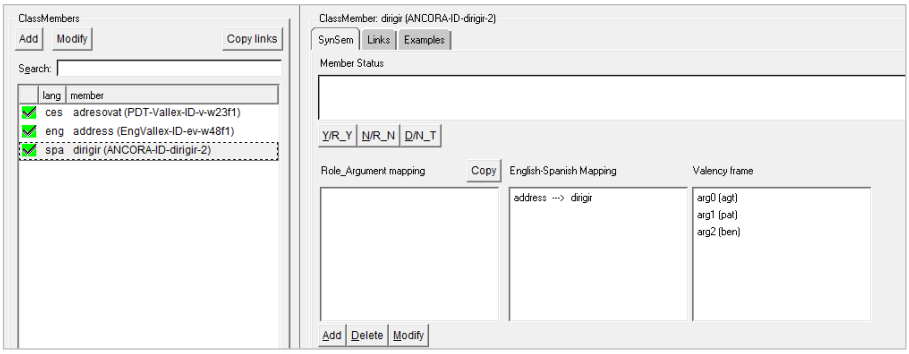

Figure 4. SynSem tab in the editor: 'Role_Argument mapping' box (left), 'English-Spanish Mapping' box (center) and 'Valency frame' box (right) for verb *dirigir* (class *adresovat/address/dirigir*)

[8] *#any* means 'either somebody or something' (as in *Jane/The winter has come*, where either a person or thing can realize the action).

[9] Some roles in the *Roleset* box are labelled as 'fn', meaning that the role takes the name from FrameNet.

[10] The language of the role definition is the language set by default during the installation in the config_file_multi (see section 2.4.3).

The valency frame of each class member with AnCora-ID (e.g., *dirigir*, class *adresovat/address/dirigir*) is displayed in the box *Valency frame* (e.g., arg0 (agt), arg1 (pat), arg2 (ben)) based on the frame described in AnCora[11]. The mapping of this particular CM to its English translational counterpart is displayed in the *English-Spanish Mapping* box (e.g., *address → dirigir*).

The content in the boxes *English-Spanish Mapping* and *Valency Frame* in not intended for editing but for providing annotators with the necessary information for the mapping of arguments to roles, specified in the left box of the S*ynSem* tab (Figure 5).

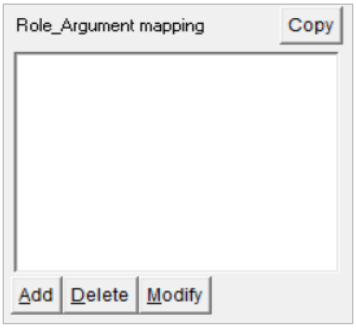

Figure 5. Role_Argument mapping box before mapping

To assign a role to an argument, click 'Add' to open the editor window (Figure 6) and select a 'Functor' from the various options displayed and the 'Role' assigned to that particular functor. Click 'OK+Next' to annotate another pair. When the mapping is complete, click 'OK'.

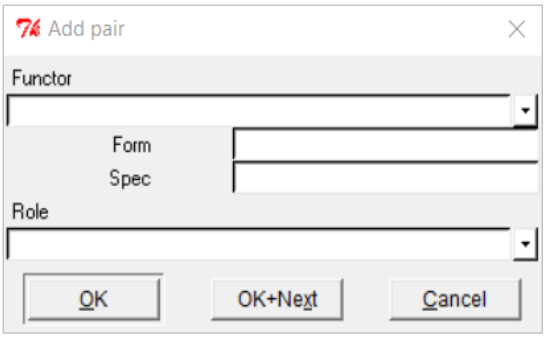

Figure 6. The editing window for adding a mapping for verb *dirigir* (class *adresovat/address/dirigir*)

If it is necessary to restrict a certain mapped valency member to a specific form, use the box 'Form' to specify such condition (Figure 7). For example, arg2 in *dirigir(se)* in the class

[11] Apart from the valency frame, AnCora also provides information regarding the lexical semantic structure of the verb sense, the syntactic realization and thematic role of each argument. It also provides examples for each sense, so it is important that annotators check the information provided by AnCora when mapping arguments to semantic roles in the editor (see section 1.8.1).

adresovat/address/dirigir is always realized by a prepositional phrase introduced by the preposition *a*:

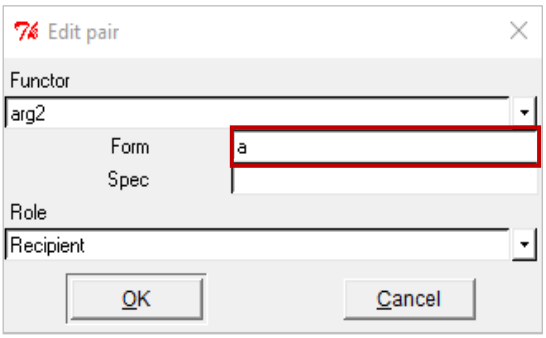

Figure 7. Role_argument mapping of the verb *dirigir* (class *adresovat/address/dirigir*), with specification of the preposition required by the verb in *arg2*

The final version of the annotation in the *Role_Argument mapping* should look like the example in Figure 8:

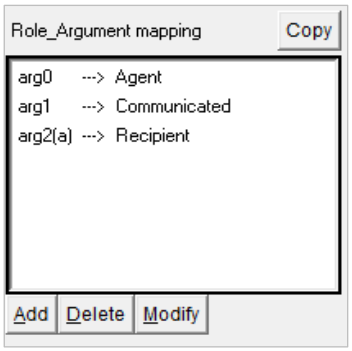

Figure 8. Example of annotation of the verb *dirigir* (class *adresovat/address/dirigir*)

In general, semantic roles are preferably mapped to valency elements using functors. However, in those cases where meaning slots cannot be included as valency slots in the valency frame, proxy functors (i.e., #any, #sb, #sth, #smh, #smt, #swh) are used instead (Figure 9).

Figure 9. Role mapping to a proxy functor

If a member of the valency frame of a potential class member cannot be really mapped to the predefined Roleset, then the class candidate does not belong to the class and should be discarded.

1.5.2 *SynSemClass-ID*

For verbs with SynSemClass-ID, no valency frame is provided by default (Figure 10). These verbs are included in case the particular meaning of the class is not included as a sense in AnCora. For example, the verb *aprobar* is included in AnCora with the sense 'aprobar una ley' ('pass a bill'), but not with the meaning of the class *absolvovat* ('A Protagonist go through an Activity'). Even if both senses share the same argument structure (i.e., arg0, arg1), they differ in the semantic roles they are mapped to.

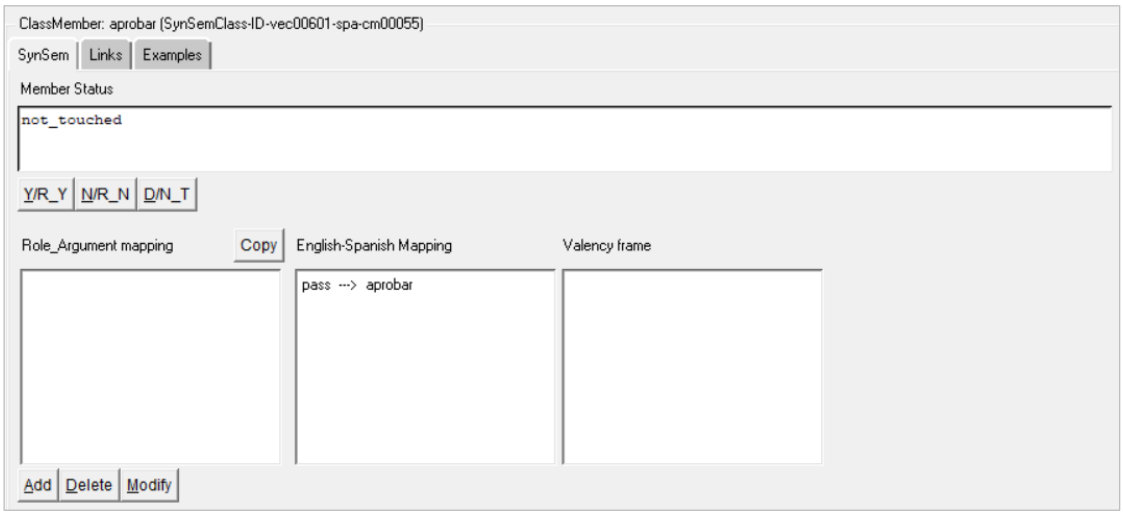

Figure 10. SynSem window for verbs with SynSemClass-ID

The annotation of verbs with SynSemClass-ID is very similar to the procedure described for the verbs listed in AnCora. To add a role-argument mapping, click 'Add' and select a functor from the list (for this type of verbs, functors are labelled as SA0, SA1, etc.) and the corresponding role (e.g., SA0 is mapped to the role 'Protagonist', while SA1 is mapped to the role 'Activity') (Figure 11).

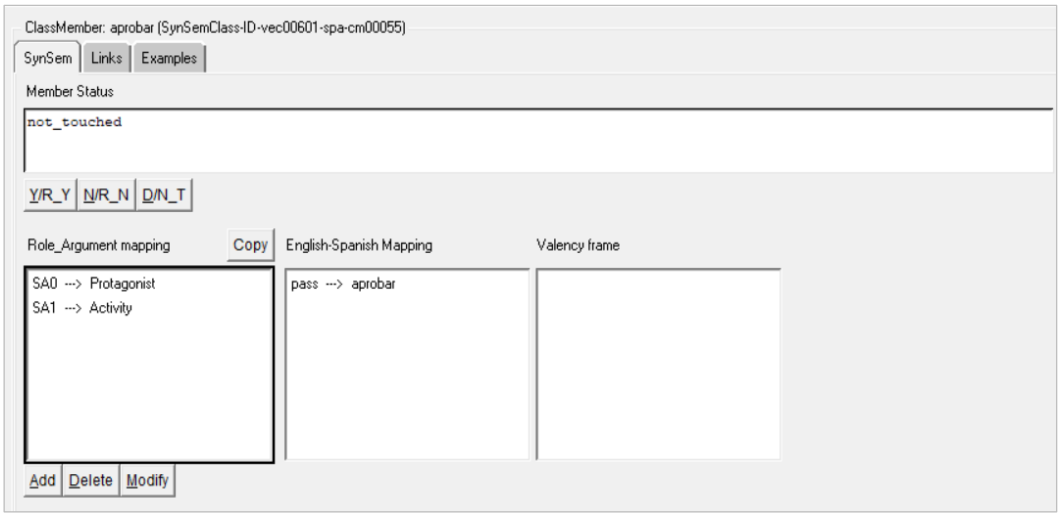

Figure 11. Role_Argument mapping for verbs with SynSemClass-ID

For consistency, the mapping of role and arguments in verbs with the SynSemClass-ID follows the mapping in AnCora, which is based on PropBank. That is, S0 is mapped to roles denoting an agent (in the narrow sense), S1 is mapped to the roles indicating an affected participant and S2 is mapped to the beneficiary or addressee of the action. Arguments S3 onwards are used with adjuncts, which may vary depending on the semantics of the verb. For example, S3 can be mapped to the role 'Area 1' to indicate the origin of the action in classes such as vec00022 (*hnout se*/*move*/*mover*), for which the Roleset defined contains the roles 'Mover', 'Area_1' and 'Area_2'. However, the same argument can be also mapped to other roles in other classes, for example, to indicate an instrument, such as 'Payment' in class vec00035 (*nakoupit*/*buy*/*comprar*/*einkaufen*), for which the Roleset defined contains the roles 'Buyer', 'Goods', 'Party_benefited', 'Payment' and 'Seller'.

1.6 Restrictions and Notes

Finally, some verbs may need to fulfil a specific condition to be considered as members of a synonymous class. Annotators can specify the necessary conditions in the box *Restrictions* as free text. This is particularly common in set phrases or idioms, pronominal uses of the verb, whether the CM is an antonym[12] or if the verb is restricted to a context. For example, for class *absolvovat*/*pass*/*aprobar* (vec00601), 'aprobar' is restricted to the context of 'pass a test, exam or course' (Figure 12).

[12] If the CM is an antonym, please specify it as 'Antonym'.

Figure 12. Example of a restriction for verb *aprobar*

The 'Restrict' box is also used to specify the obligatory phrase that precedes the verb in LVCs, idioms or in other set expression. For example, for *hacer erupción* ('erupt'), the NP *erupción* should be specified as a restriction. For ease of identification, annotators should also write 'LVC' in the box 'Member note' (Figure 13).

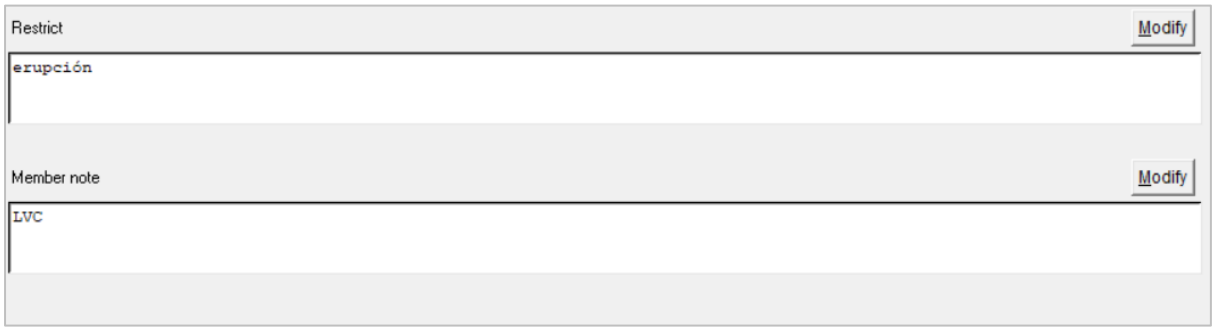

Figure 13. Notes for a LVC

Other aspects that the annotator may consider relevant for each individual CM can be also added in the 'Member note' box, whether the CM has negative/positive connotations unique to that particular verb, whether it is restricted to a particular register (e.g., colloquial), more frequent in a particular dialectal variety (i.e., Latin American/Peninsular Spanish) or any other aspect annotators consider to be relevant (Figure 14):

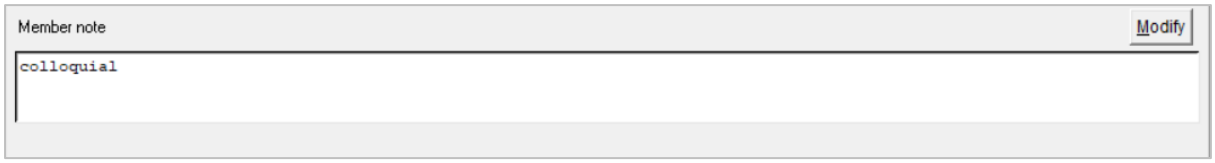

Figure 14. Example of a note for the member of a class

1.7 Spanish class name

Each synonym class in the editor is named using the verb in each language that is considered to be the most prototypical meaning (if available). In order to facilitate the search by class name in Spanish, a Spanish Class Name is assigned to every class containing a Spanish CM. To add a name for a class, select the verb with the most prototypical meaning that may be a good candidate to be the class name from the *ClassMembers* list (shaded in grey) (Figure 15) and then click the 'Set' button (Figure 16).

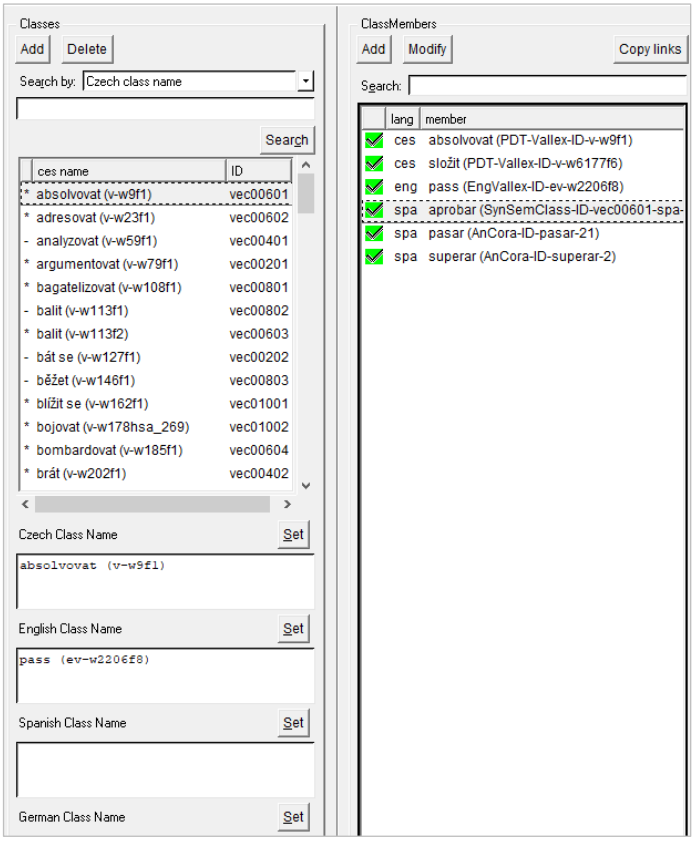

Figure 15. Candidate verb from the *ClassMembers* list to be class name

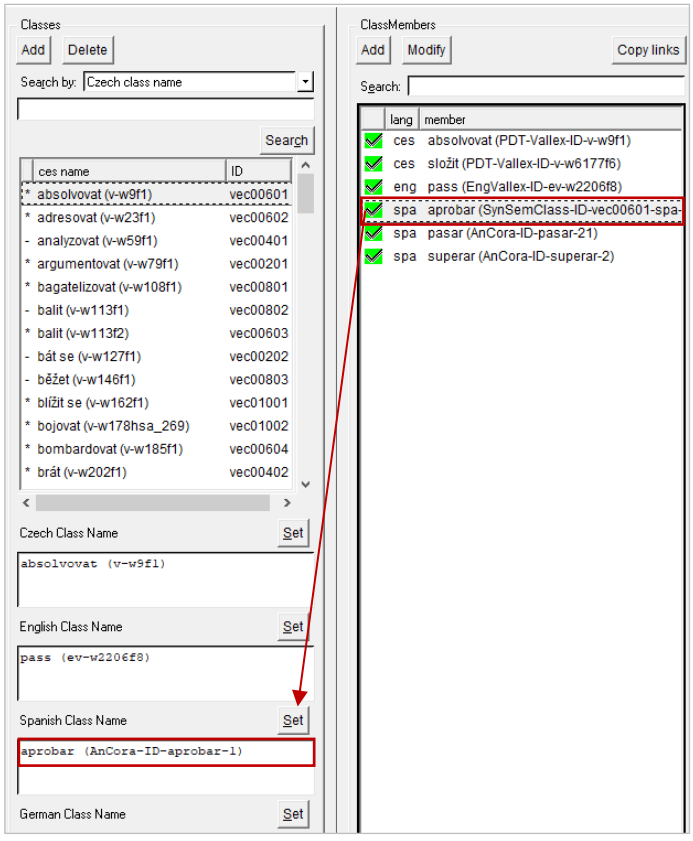

Figure 16. Class name set

Figure 17 shows an example of a CM annotated. The class has been assigned a Spanish name (1) and the particular verb has been marked as belonging to the class (2), the roles have been mapped to arguments (3), and the restrictions of use and relevant notes have been added (4).

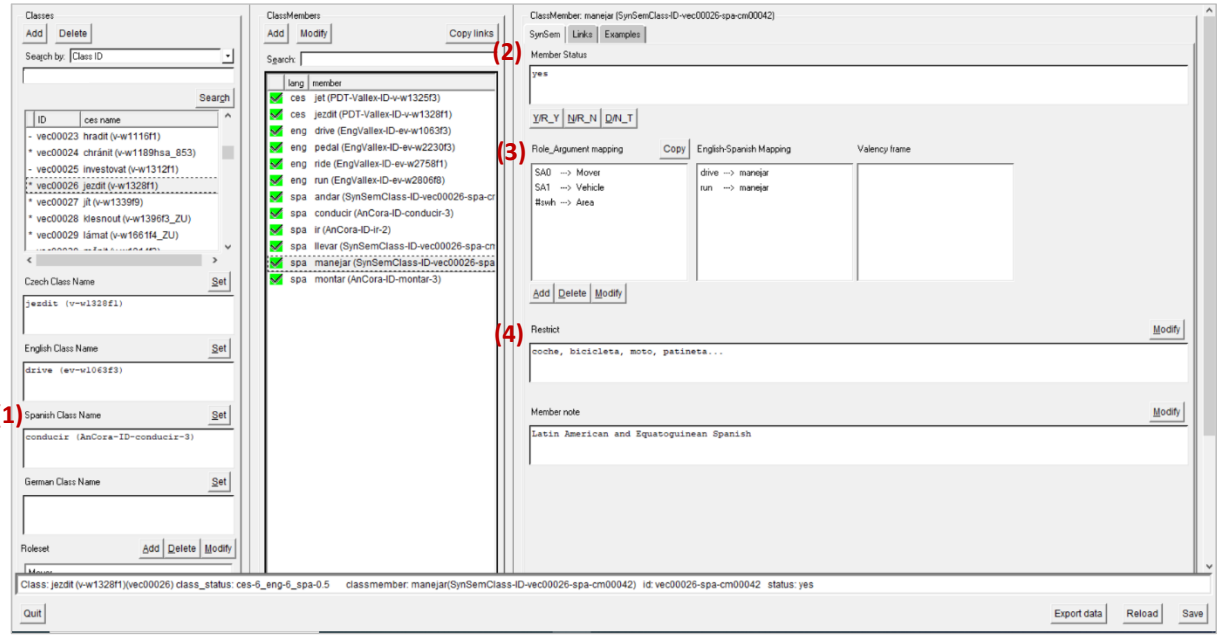

Figure 17. Complete Role_Argument mapping for class *jezdít*/*drive*/*conducir* (vec00026)

1.8 Links to other lexical resources

Linking the synonymous verbs in the lexicon to other online lexical resources has two aims:

i) to allow interoperability, thus enriching the tool for its use in computer linguistics, and

ii) to gain a better insight of the individual meanings of the verbs included and of their characteristics.

The links for each of the external resources used for Spanish are added in the *Links* tab (Figure 18). The same CM in the lexicon can be linked to one or more than one resource and it is also possible that no links are available for some CMs.

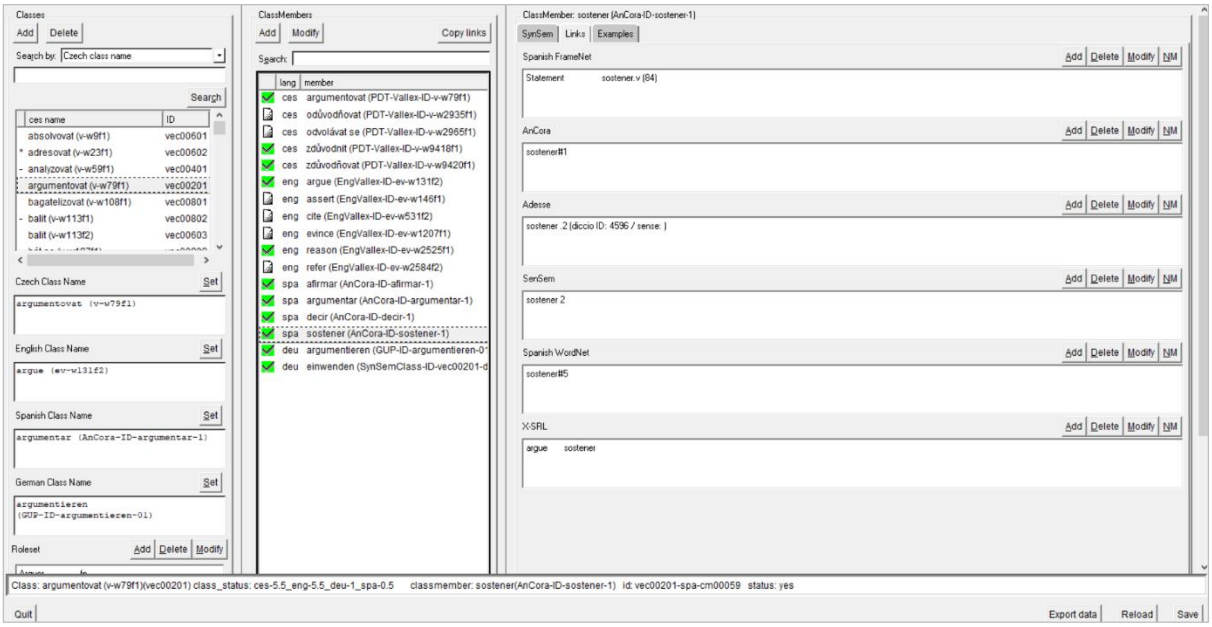

Figure 18. Links in the SynEd editor for the CM *sostener* (class *argumentovat/argue/argumentar*)

The resources used for Spanish data and the procedure to include links to each one are described in more detail in Sections 1.8.1 to 1.8.5.

1.8.1 AnCoraVerb_ES Lexicon

Of the three verbal lexicons included in the editor, the AnCoraverb_Es Lexicon (henceforth, AnCora)[13] is the main source for the identification of senses for Spanish. Each Spanish verb is identified with an ID, which includes the name of the resource (e.g. AnCora), the verb (e.g., *argumentar*) and the sense number in the online version of the lexicon (e.g., 1).

The advantages of using AnCora as the main resource for sense identification are:

(1) It provides information regarding the diatheses of the verb sense (e.g., *default*, *impersonal*, *passive*, etc.).

(2) It defines a lexical semantic structure (LSS) for each verb sense based on the argument structure described for that particular sense.

(3) It provides links to other lexical resources in English (under the heading *AnCoraNet*) (namely, PropBank, VerbNet, FrameNet and WordNet) which are also used for the annotation of English synonymous verbs in our lexicon. Although discrepancies may exist, annotators may find useful to compare the links provided by AnCora to those selected for the English members of a class in our lexicon.

[13] http://clic.ub.edu/corpus/es/ancoraverb_es

(4) It provides the argument structure of the verb (*Argument*), but also the syntactic realization (*Function*) and the thematic role (*Theme*) of each argument.[14]

(5) For each sense, AnCora provides a list of examples that conform to the argument structure defined for that particular sense.

Figure 19 shows an example of a verb sense in AnCora:

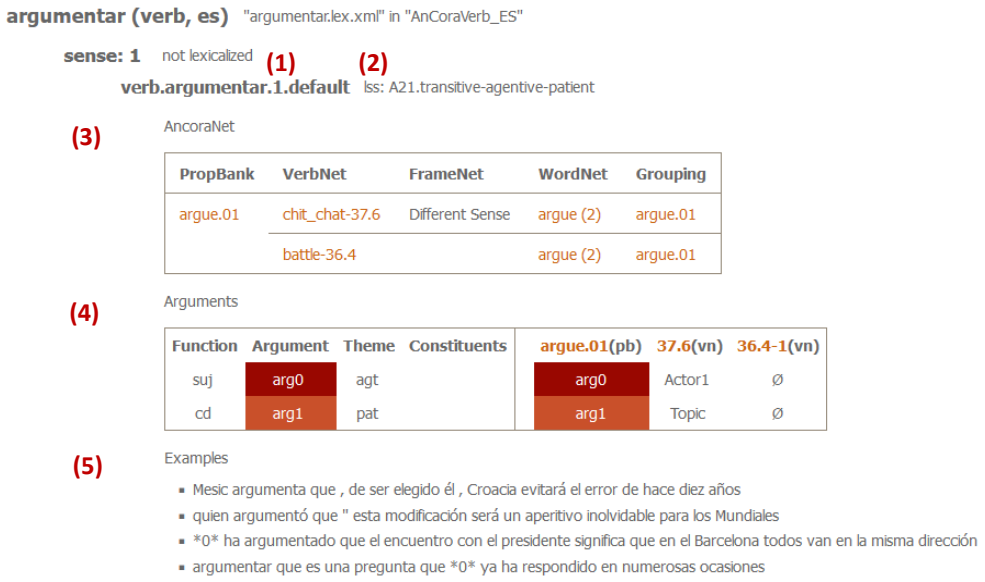

Figure 19. Information for the verb *argumentar* displayed in AnCora

As AnCora is used as the main resource, links are added automatically (Figure 20).

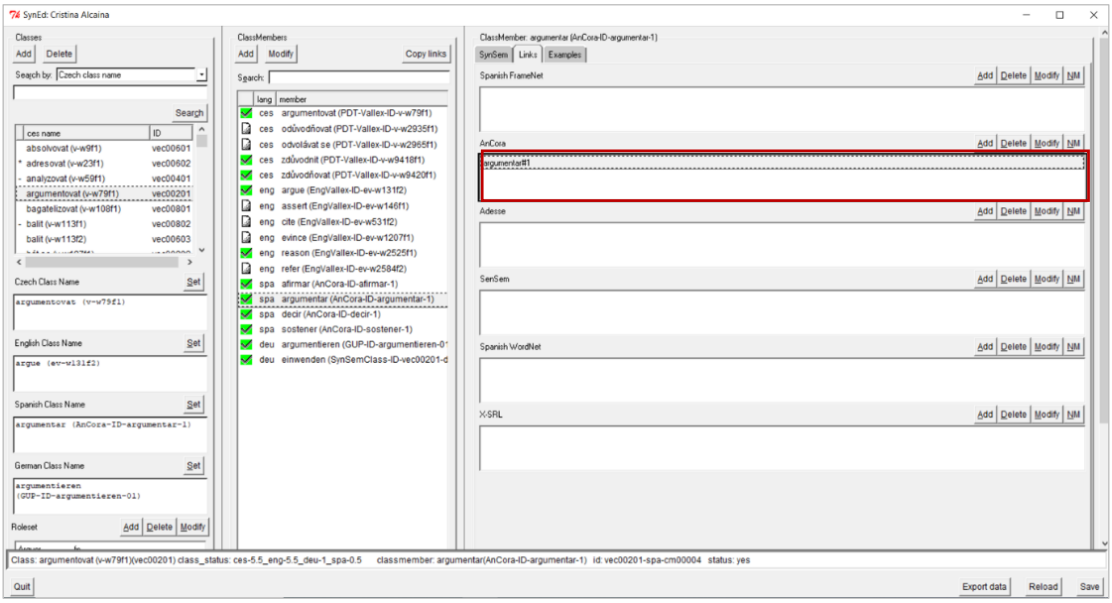

Figure 20. Links tab for the CM *argumentar*

[14] More information on arguments, functions and thematic roles in AnCora can be consulted here: http://clic.ub.edu/corpus/webfm_send/49.

For verbs with AnCora-ID, links to AnCora are automatically added. The online version of the lexicon can be accessed from the editor by double-clicking.

1.8.1.1 Alternations

For each verb in AnCora, there is information regarding the number of senses, the alternation (i.e., whether *default*, *passive*, *impersonal*, see below), links to synonymous senses in other English lexical resources (e.g., VerbNet), information regarding the argument structure of the sense and a list of examples (if available). In this regard and as mentioned above, it is important to note that AnCora distinguishes several 'variants' (alternations) of the same sense (i.e., default, passive, semelfactive, anticausative, etc.). Each particular type is marked with a label after the number sense (e.g., *verb.decir.1.**default***, *verb.decir.1.**passive***, *verb.decir.1.**impersonal***). At the moment, the automatic argument structure provided in SynEd is based on the sense labelled as 'default' in AnCora, but it is possible that the structure of a particular class in SynSemClass matches one of the alternations. In these cases, a note should be added to specify that the argument structure provided belongs in fact to one of the alternations listed in AnCora. For example, the defined Roleset for class vec00012 in SynSemClass is *Authority*, *Permitted*, *Permitte*. For the CM *dejar*-AnCora-ID-dejar-3, the argument structure provided by default is based on the sense verb.dejar.3.default in AnCora (arg0 (agt), arg1(pat), argm(adv)). However, it the alternation verb.dejar.3.benefactive (arg0(agt), arg1(pat), arg2(ben)) the one expressing the meaning of the class.

In these cases, annotators should add the functors contained in the valency frame of the alternation ('arg2' in our example) in the Role_Argument mapping and a note specifying this in the 'Member note' box as in Figure 21.

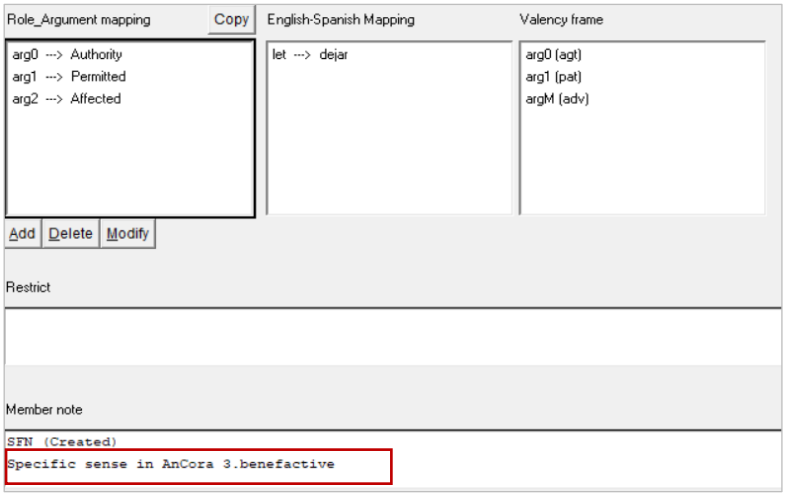

Figure 21. Annotation of a CM whose argument structure matches the argument structure of an alternation in AnCora

1.8.1.2 Lexical Semantic Structure (LSS)

In AnCora, arguments labelled as arg0 and arg1 are considered to be obligatory, but the status of arg2 or arg3 as obligatory or optional is not so straightforward (Palmer et al. 2005: 76)[15]. In AnCora, the LSS 'restricts the set of all possible diatheses it can incur in, and each verb sense is associated to one LSS' (Taulé et al. 2011: 4)[16]. Therefore, especially in cases where the argument structure defined for a particular verb sense in AnCora contains three or more arguments, annotators can use the LSS classification of that particular sense as a guide[17]. Sections 1.8.1.2.1 and 1.8.1.2.2 illustrate this with two specific examples.

1.8.1.2.1 *Arg2 is obligatory: dirigir (sense 2)*

The LSS described for *dirigir* (sense 2) in AnCora is *A32.ditransitive-patient-benefactive*, which involves the presence of three arguments: arg0(agt), arg1(pat) and arg2(ben), the latter realized by a prepositional complement (Figure 22):

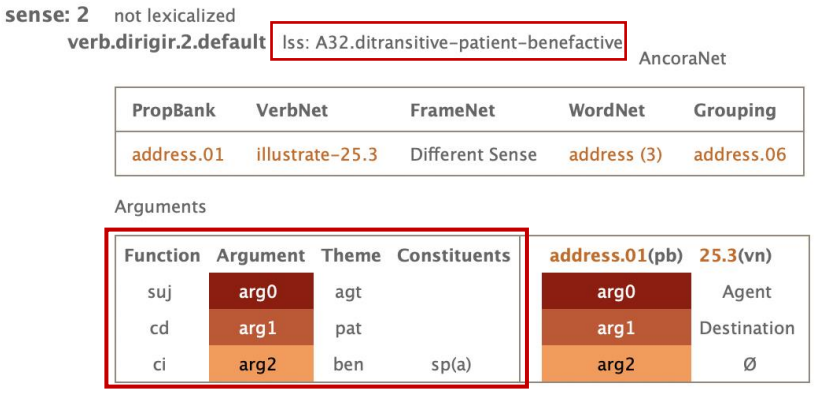

Figure 22. Argument structure of *dirigir* in AnCora

Therefore, arg0, arg1 and arg2 need to be obligatorily mapped to each of the roles in the Roleset for the class where *dirigir* is a CM (class *adresovat*/*address*/*dirigir*). Specifically, ar0 (Agent), arg1 (Communicated), arg2 (Recipient) (Figure 8 above).

1.8.1.2.2 *Arg2 is optional: abrir*

In contrast, the LSS of *abrir* is *A11.transitive-causative*, which in this case involves the presence of two arguments: arg0(cau) and arg1(tem). That is, of all the arguments described

[15] Palmer, Martha, Gildea, Daniel & Kingsbury, Paul. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1): 71–106. https://doi.org/10.1162/0891201053630264

[16] Taulé, Mariona, Martí, M. Antònia. & Borrega, Oriol. 2011. *AnCora 2.0: Argument Structure Guidelines for Catalan and Spanish*. Available at: http://clic.ub.edu/corpus/webfm_send/52.

[17] See Appendix 1 for a summary table with the arguments corresponding to each LSS in AnCora.

for *abrir* (Figure 23), only arg0 and arg1 need to be considered for the mapping with the roleset while arg2, and argM are optional and, thus, they do not need to be mapped to a role.

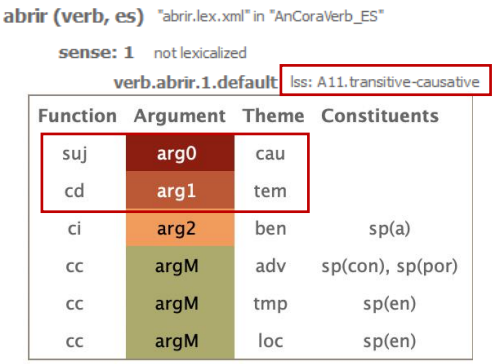

Figure 23. Argument structure of *abrir* in AnCora

1.8.1.3 Light Verb Constructions (LVCs) in AnCora

AnCora includes the argument labelled as 'argL' for lexicalized complements of light verbs, as in *hacer fuego* ('make a fire') in Figure 24.

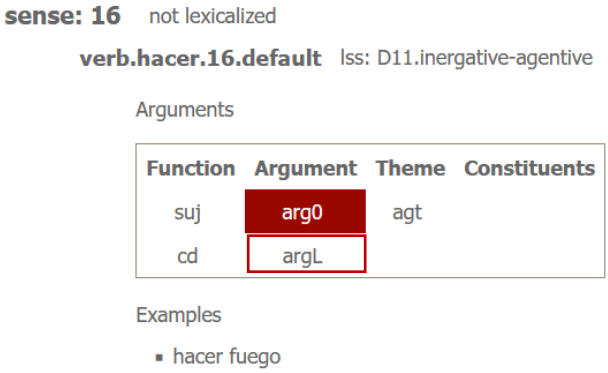

Figure 24. Argument argL in AnCora

For example, for the CM *hacer erupción* ('erupt') in vec00018, the argument argL would refer to 'erupción', which will be specified for that particular member in the editor by adding a note in the 'Member note' box specifying that this is a LVC (see Figure 13).

1.8.2 SenSem Corpus Verbal Lexicon

The list of verbs contained in the SenSem Corpus Verbal Lexicon (henceforth, SenSem)[18] is restricted to ca. 260 units, thus it is possible that some of the Spanish verbs annotated are not available in this resource. For each candidate sense, it is necessary to specify the verb and sense to which it refers in SenSem.

[18] http://grial.edu.es/sensem/lexico/main

To add a link to SenSem, click ‘Add option’ and the editor will display the window in Figure 25:

Figure 25. Add SenSem link window

Click ‘Search’ and the editor will open the SenSem search page. Choose a verb from the drop-down list on the left (Figure 26):

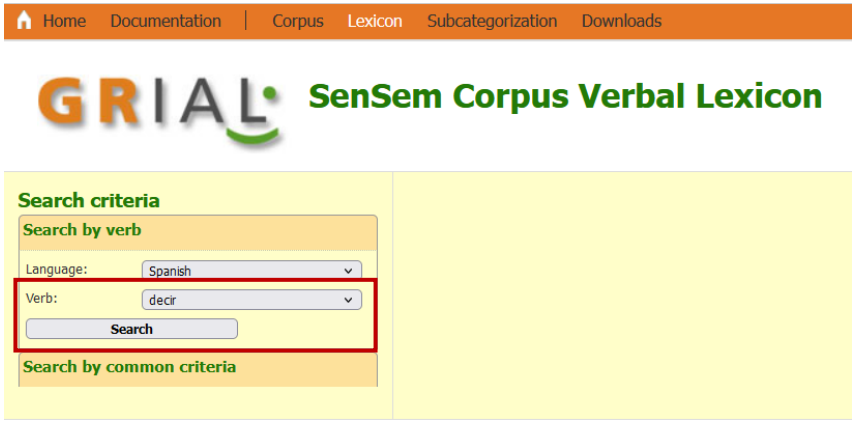

Figure 26. SenSem search page

It will open a page with the available senses for that verb (Figure 27). Annotators should go through the senses displayed and choose the one that better captures the meaning of the class of the member:

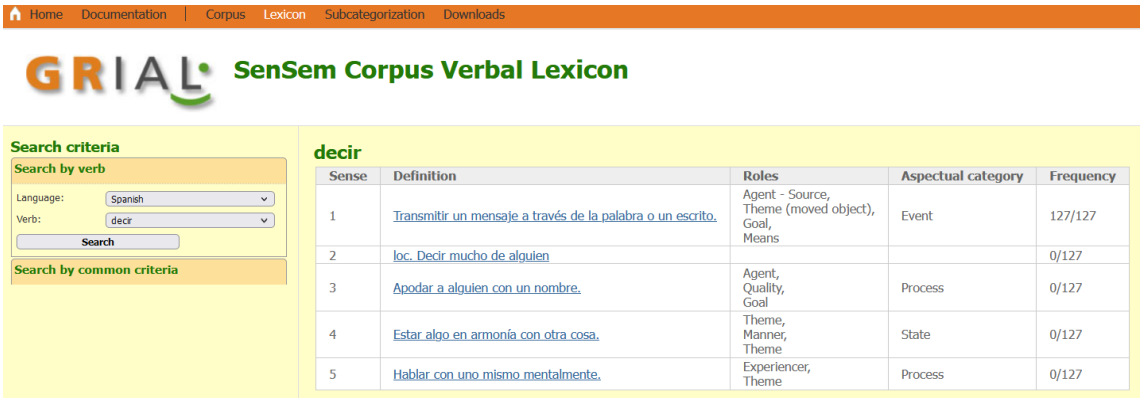

Figure 27. List of senses for the verb *decir* in SenSem

Once the appropriate sense has been identified, annotators must consider the sense ID in the URL (‘sense=’). For example, CM candidate sense ‘transmitir un mensaje a través de la palabra o el escrito’ of the verb *decir*, the sense ID is 1 (‘sense=1’ in the URL) (Figure 28).

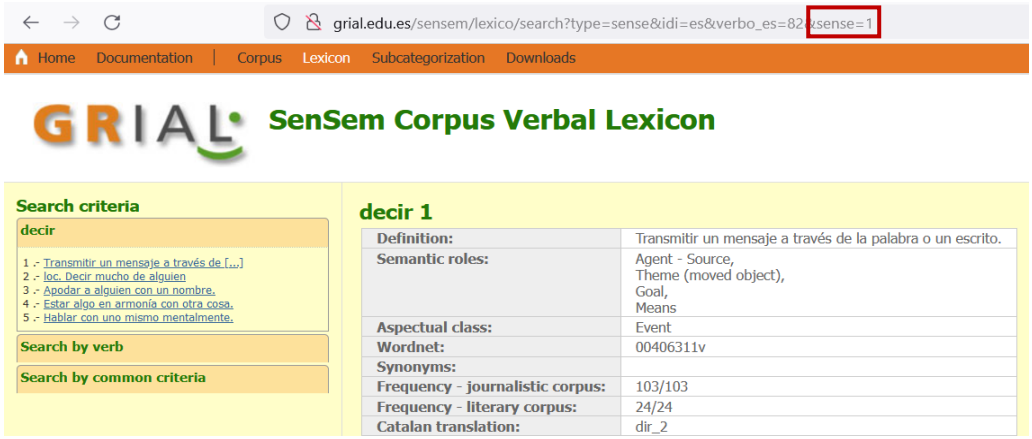

Figure 28. Verb and sense ids for *abrir* in URL

The ID in the URL needs then to be filled in the box 'Sense' in the editor (Figure 29):

Figure 29. Add SenSem link window

If none of the senses in SenSem refers to the meaning of the class, click 'NM'.

1.8.3 ADESSE

The third verbal lexicon to which verb senses in *SynEd* are linked is ADESSE[19]. The lexicon contains information regarding semantic roles and valency structure ('Clasificación semántica y potencial valencial'), a syntactic-semantic scheme ('Realizaciones valenciales (Esquemas sintáctico-semánticos)', which provides information regarding the different argument structure depending on the voice) and a list of examples with detailed information regarding their argument structure.

To add a link to ADESSE in SynEd, click 'Add' and the editor will display the window in Figure 30:

[19] http://adesse.uvigo.es/data/

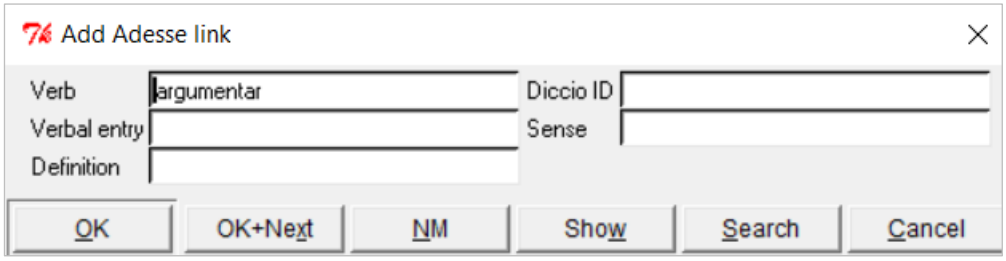

Figure 30. Add ADESSE link for the verb *argumentar*

Click 'Search' and the editor will open the ADESSE search page (Figure 31):

Figure 31. ADESSE search page

A note of caution before moving on to the process to add a link to ADESSE. Apart from senses (identified by numerals), ADESSE can also distinguish various verbal entries (identified by Roman numerals) for the same verbal lemma. These verbal entries can be at the same time divided into further senses. Therefore, when adding a link to ADESSE, three possible scenarios arise. These are described in more detail in sections 1.8.3.1 to 1.8.3.3.

1.8.3.1 *Argumentar* verbs

The lemma has only one sense. For this type of verbs, it is only necessary to fill in the box 'Sense' with the ID provided in the URL (Figure 32).

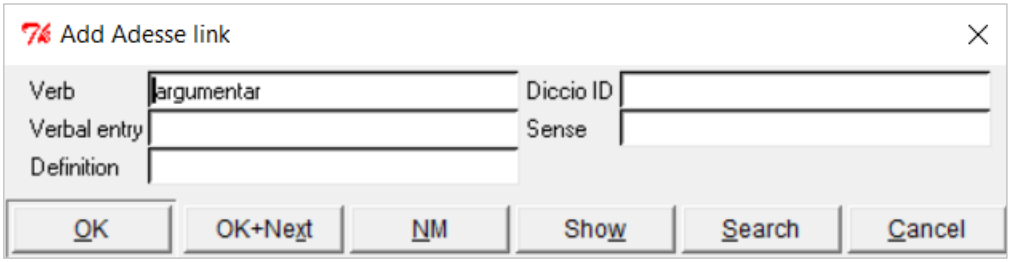

Figure 32. Add ADESSE link for *argumentar*

ADESSE displays the following information for each verb (Figure 33):

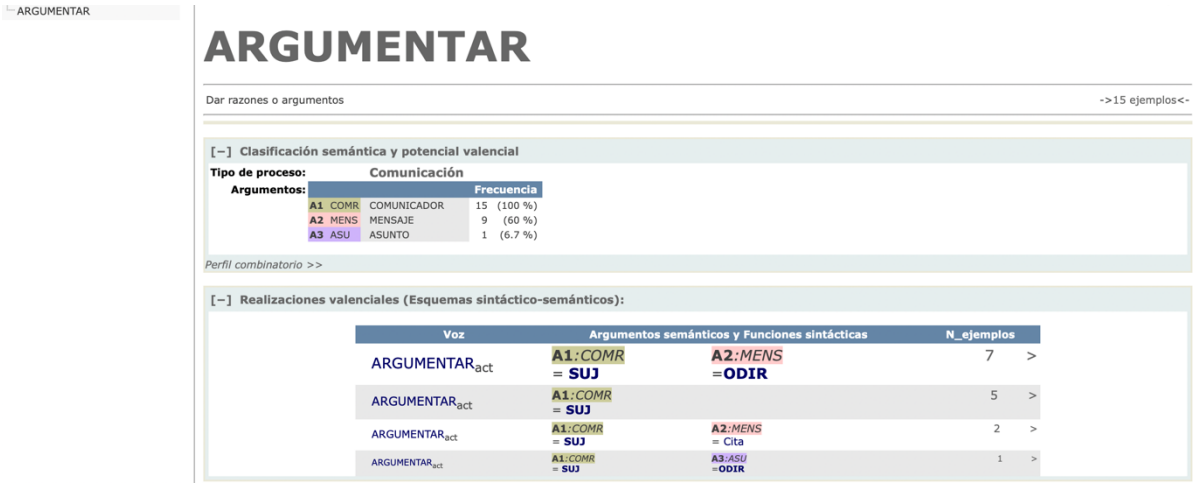

Figure 33. Verb *argumentar* in ADESSE

It is important to keep in mind that agentive entities are treated slightly different in ADESSE and in AnCora. Specifically, AnCora uses *arg0* for agents, causes and perceivers (i.e., subjects) and *arg1* for patients and themes (i.e., objects). In ADESSE both A0 and A1 are reserved for entities realizing the syntactic function of subject but the use of one or the other captures a semantical difference: A0 is restricted to the agents of causative events, while perceivers, non-causative agents (such as Protagonists) or affected entities are labelled as A1.

Figure 34 compares the argument structure of the verb *argumentar* in AnCora and ADESSE:

Figure 34. Comparison of argument structure in AnCora (left) and ADESSE (right) for verb *argumentar*. The argument arg0 (agt) in AnCora corresponds to A1 (COMPR) in ADESSE

If annotators consider that the sense in ADESSE matches with the meaning of the class, the next step is to access the ID of the verb in ADESSE. To do this, annotators must click on the verb in small capitals on the left side of the website (Figure 35).

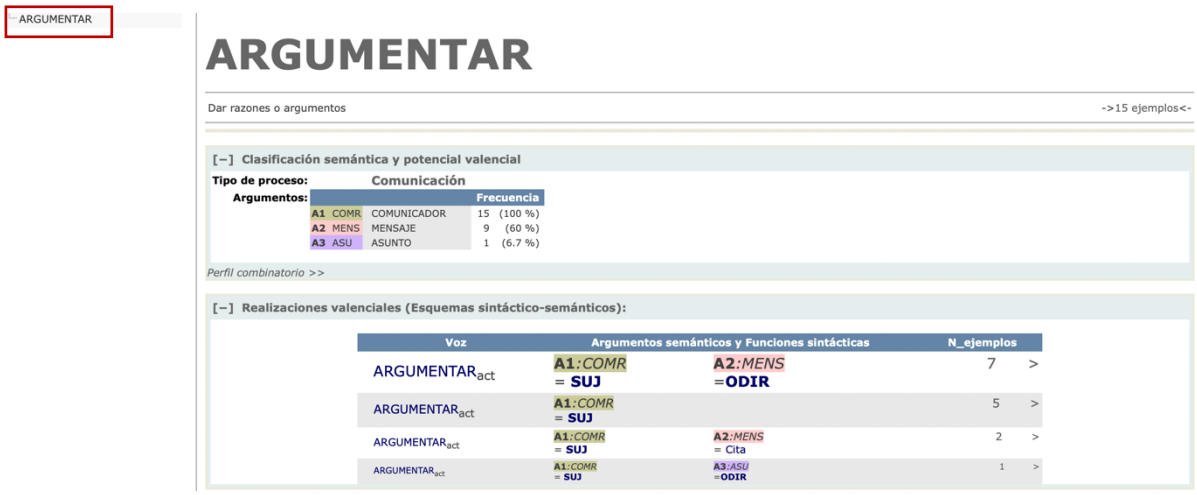

Figure 35. Argument structure for *argumentar* in ADESSE

The ID of the verb is found in the URL, identified as 'Sense' (Figure 36). For *argumentar*, the ID of the verb in ADESSE is 315:

Figure 36. 'Sense' in ADESSE for *argumentar*

This is the number that needs to be introduced in our editor. In this type of verbs, the boxes 'Verbal entry' and 'Definition' are left in blank (Figure 37):

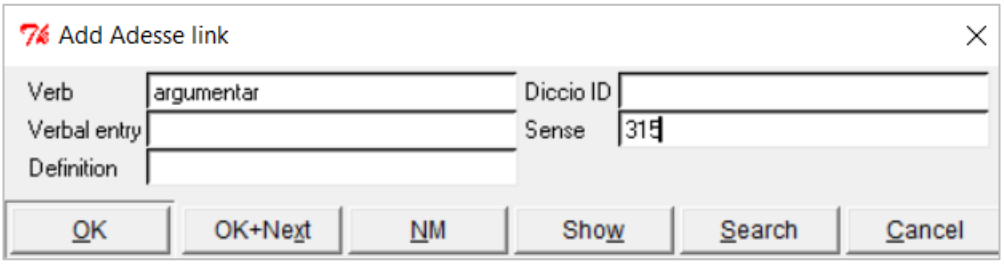

Figure 37. Required information for adding a link to ADESSE (*Argumentar* verbs)

1.8.3.2 *Superar* verbs

One lemma has various senses (identified by numerals) and/or subsenses (identified by a numeral followed by a letter).

Search for the verb in ADESSE and go through the senses and the information provided. For example, for the verb *superar* (Figure 38), ADESSE distinguishes three senses (on the left). Senses 1 and 2 are further divided into subsenses.

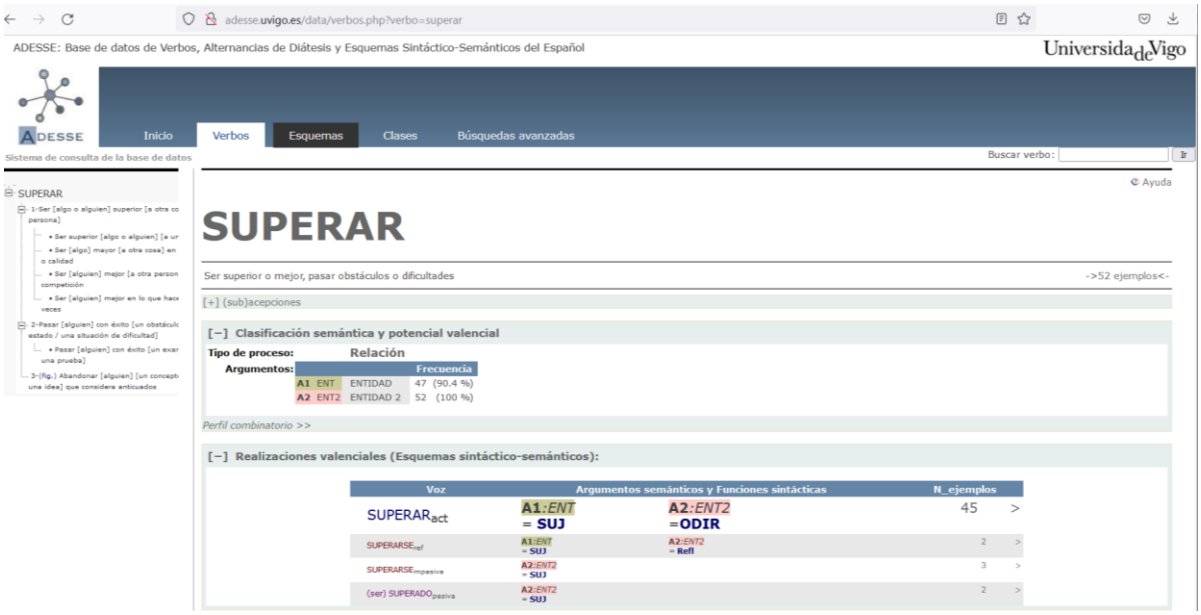

Figure 38. Argument structure for *superar* in ADESSE

The meaning of class 'aprobar-AnCora-ID-aprobar-1' is captured by sense 2a in ADESSE ('Pasar [alguien] con éxito [un examen o prueba]') (Figure 39). The ID of the verb ('diccio_id=6283') is provided in the URL (red box):

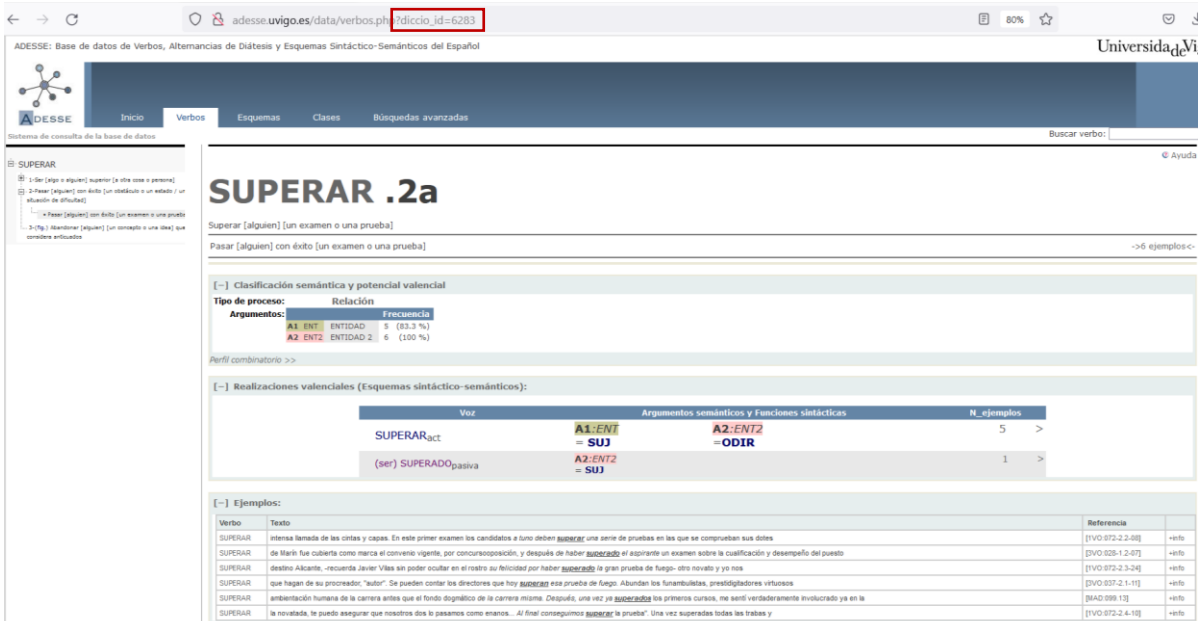

Figure 39. Sense 2a of *superar* in ADESSE

To add the link in the editor, fill in the fields in the *Add Adesse link* window that correspond to *Definition* (e.g., 2a) and *Diccio ID* (e.g., 6283), as shown in Figure 40. The field *Verbal entry* is left blank.

Figure 40. Required information for adding a link to ADESSE (*Superar* verbs)

1.8.3.3 *Volver* verbs

One lemma is divided into various verbal entries (identified by Roman numerals in ADESSE).

1.8.3.3.1 *Type 1*

If the verbal entry is not further divided into senses (e.g., *volver III*), proceed as in section 1.8.3.1 (*argumentar* verbs) and use the number identified as 'Sense' in the URL (Figure 41 and 42).

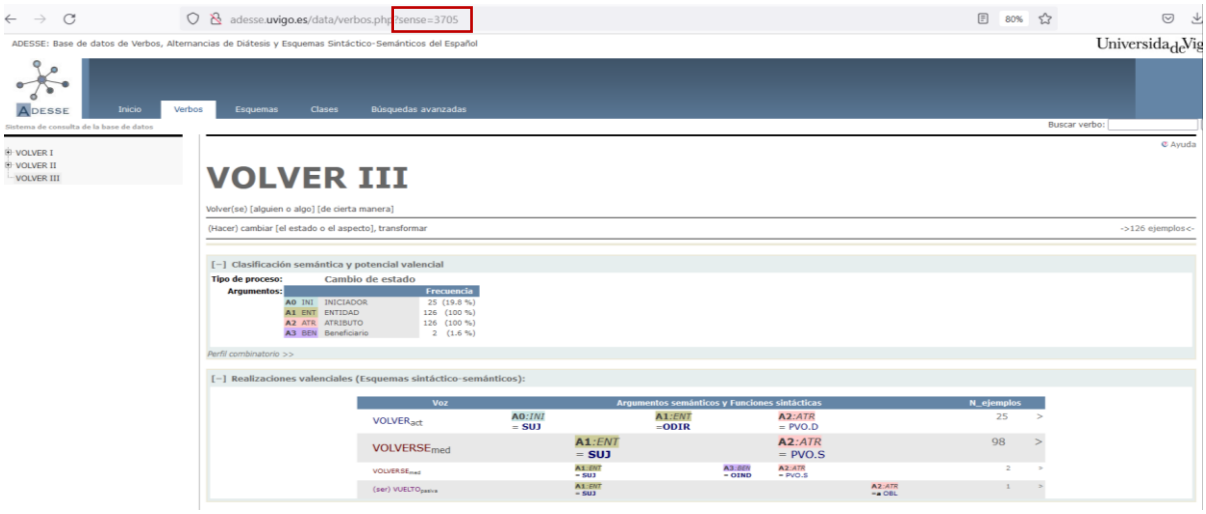

Figure 41. Verbal entry *Volver III* in ADESSE

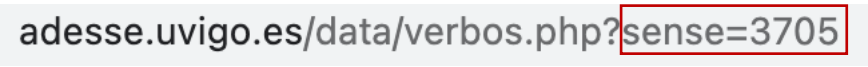

Figure 42. 'Sense' in ADESSE for *volver III*

Then go to SynEd and fill in the field *Verbal entry* with the Roman numeral for that entry in ADESSE (e.g., III) and the field 'Sense' with the ID in the URL (e.g., 3705) and leave the box *Definition* blank (Figure 43).

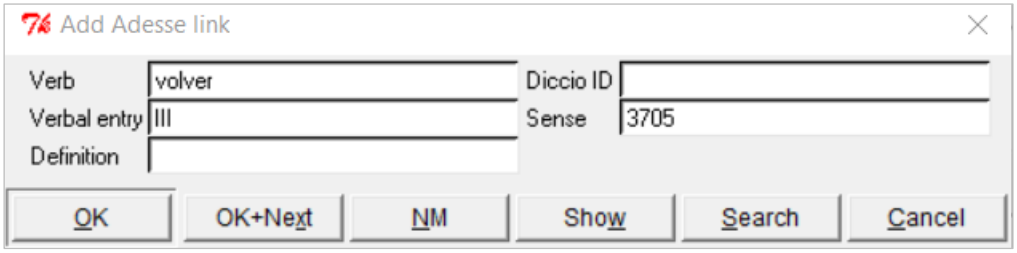

Figure 43. Required information for adding a link to ADESSE (*Volver III* verbs)

1.8.3.3.2 *Type 2*

If the verbal entry is further divided into senses (e.g., *Volver I*), go through the senses and choose the one that better captures the meaning of the class and go to the URL to find the 'diccio_id' (Figure 44):

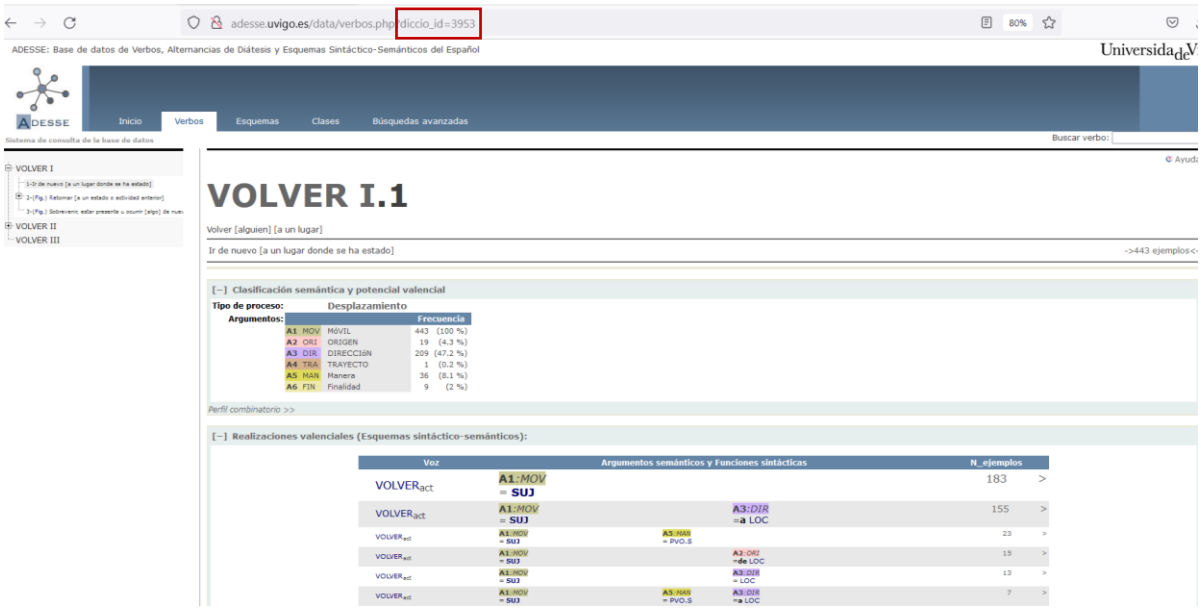

Figure 44. Sense 1 for verbal entry *volver I* in ADESSE

Now go to SynEd and fill in the fields *Diccio ID* (3953), *Verbal entry* (I) and *Definition* (1) in the *Add Adesse link* window (Figure 45):

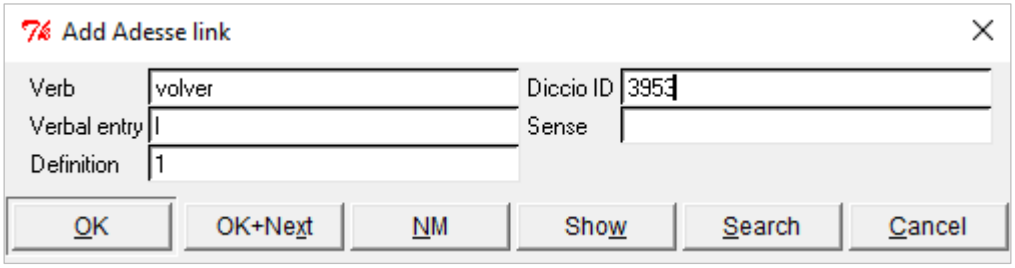

Figure 45. Required information for adding a link to ADESSE (*Volver I* verbs)

1.8.4 Spanish FrameNet

From the point of view of semantic roles, the most important source and inspiration for the *SynSemClass* lexicon is *FrameNet*[20], which uses semantic roles (i.e., Frame Elements, FEs) within semantic frames.

For Spanish data, the information included in our editor is based on the Spanish version of *FrameNet*, *Spanish FrameNet*[21]. *Spanish FrameNet* allows searches by both frame[22] and lexical unit[23]. Both frames and lexical units can be also searched by using the search box provided here[24].

The task of the annotator is to check whether the frame provided for the Spanish CM candidate is the same as that already selected for the English CM in the editor. It is up to the annotator to choose the frame that better captures the sense of the Spanish CM. To do this, the annotator can check the information provided by the *Lexical Entry Report* and the *Annotation Report* for each entry. Specifically:

i) The Lexical Entry Report shows a definition of the unit, the Frame to which it belongs (if available)[25], the list of frame elements and their realizations and a list of valency patterns.

ii) The Annotation Report provides information on the frame elements and annotated examples.

To add a link to Spanish FrameNet, click *Search* on the Add Spanish FrameNet link window in SynEd (Figure 46). It will open the Spanish FrameNet search page (Figure 47):

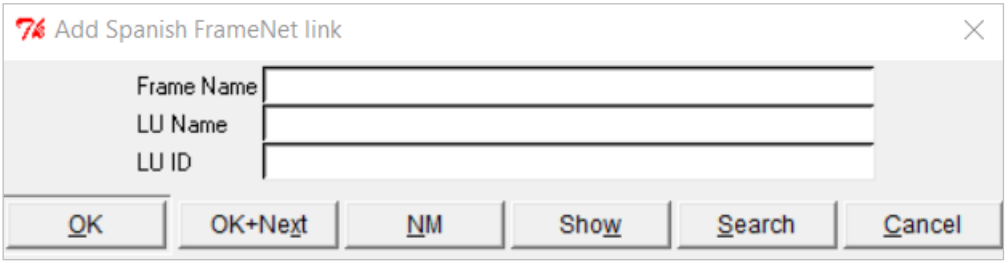

Figure 46. Add Spanish FrameNet link in SynEd

[20] https://framenet.icsi.berkeley.edu/fndrupal/
[21] http://spanishfn.org/
[22] http://sfn.spanishfn.org/frameIndex.php
[23] http://sfn.spanishfn.org/luIndex.php
[24] http://sfn.spanishfn.org/SFNreports.php
[25] See footnote 17 for frame 'Limbo'

Figure 47. Search page in Spanish FrameNet

For example, for the verb *sostener*, Spanish FrameNet returns one result (Figure 48):

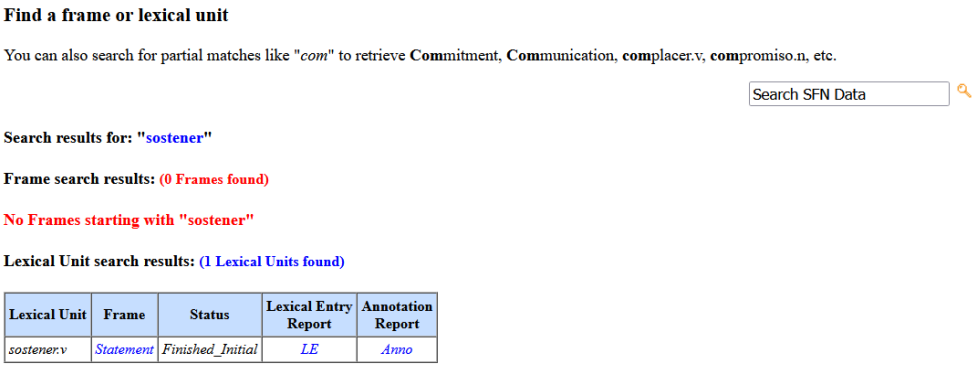

Figure 48. Overview for *sostener* in Spanish FrameNet

To access the relevant information for adding a link, annotators must open the 'Lexical Entry Report' on a new tab (by right clicking on the hyperlink 'LE' in Figure 48). The new tab will show the information displayed in Figure 49. The relevant information annotators must consider is:

i) the lexical entry (le) ID, which can be found in the URL (e.g., for the verb *sostener*, the ID is 84),

ii) the exact form of the verb (including the part of speech, as provided by FrameNet) (e.g., *sostener.v*), and

iii) the name of the frame that captures the particular sense of the verb (e.g., *Statement*).

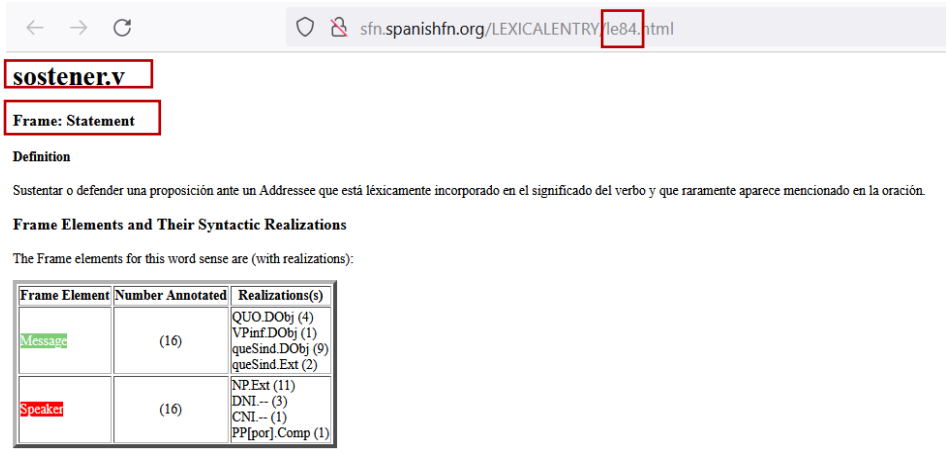

Figure 49. Information for the lexical entry *sostener* in Spanish FrameNet

Figure 50 shows the how it should look like the information for Spanish FrameNet in our editor:

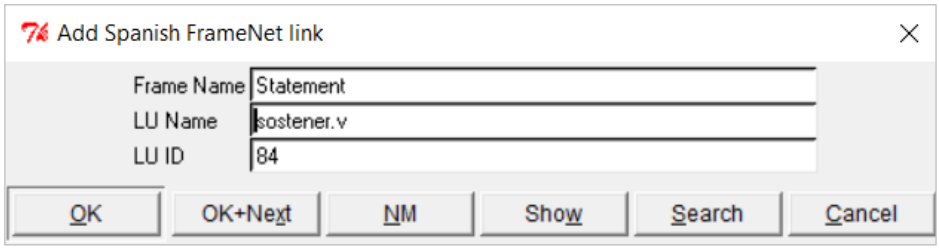

Figure 50. Required information for adding a link to Spanish FrameNet

Since the Spanish FrameNet is an ongoing project, new lexical entries and frames can be added in the future or the information provided for some of them may be refined. Entries for which complete information is not available at the moment have the status 'Created'. For example, the lexical entry *aprobar* is recorded in FrameNet within the frame *Grant_permission* (Figure 51). It is up to the annotator to consider whether the information provided is sufficient to add a link in the editor. If this is the case, the annotators must fill in the information in the editor with the name of the lexical unit (e.g., *aprobar.v*) and the name of the frame (e.g., *Grant_permission*). To access the ID of the lexical entry, annotators must open the 'Lexical Report Entry' on a new tab (Figure 52):

Search results for: "aprobar"

Frame search results: (0 Frames found)

No Frames starting with "aprobar"

Lexical Unit search results: (1 Lexical Units found)

Lexical Unit	Frame	Status	Lexical Entry Report	Annotation Report
aprobar.v	*Grant_permission*	*Created*	*LE*	*Anno*

Figure 51. Overview for *aprobar* in Spanish FrameNet

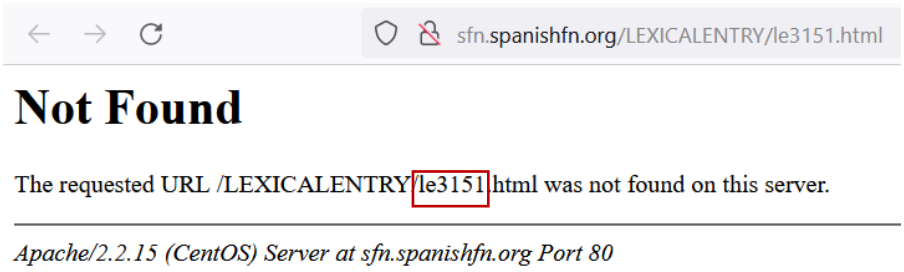

Figure 52. Error page in Spanish FrameNet for *aprobar* (frame: *Grant_permission*)

Figure 53 shows how the information will look like in the editor:

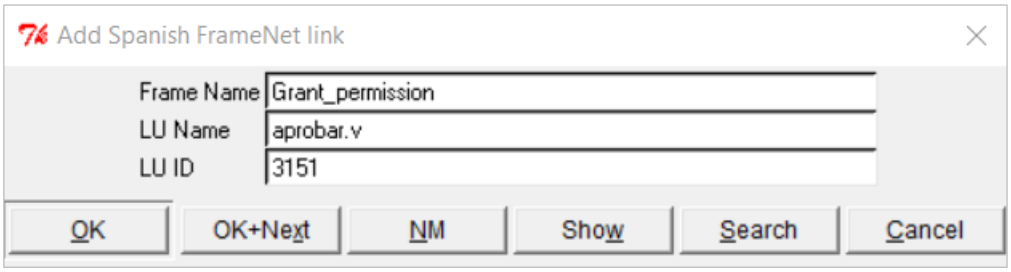

Figure 53. Required information for adding a link to Spanish FrameNet

In cases like this and in order to identify them in future versions of the editor, annotators should write 'SFN Created' as a note in the 'Member note' box (Figure 54):

Figure 54. SFN (Created) in 'Member note' box

If the verb is included Spanish FrameNet but it is not clear whether it may correspond to the meaning of the class because of lack of information, annotators have the option of fill in the

fields *Frame Name* and *LU Name* with the information available and use 'NA' for the field *LU ID* (Figure 55):

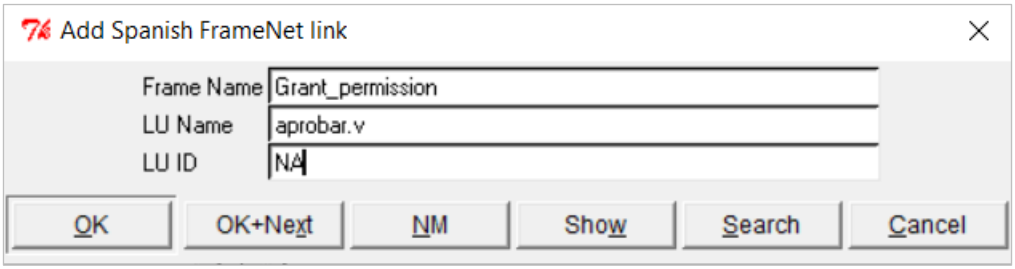

Figure 55. Required information for adding a link to Spanish FrameNet

If no mapping is available at the moment, annotators must choose the option 'NM'.

1.8.5 Spanish WordNet

The third version of the *Spanish WordNet* can be accessed using the interface provided by the *Multilingual Central Repository*[26], together with other languages such as Basque, Galician, Catalan, Portuguese and English.

To search for a specific CM, type the verb in the box (Figure 56). In the line below, choose 'Verbs' in the second drop-down list and 'Spanish_3.0' in the third drop-down list. Check the boxes 'English_3.0 ('in light blue) and 'Spanish_3.0' (in green) in the column where all the available languages appear.

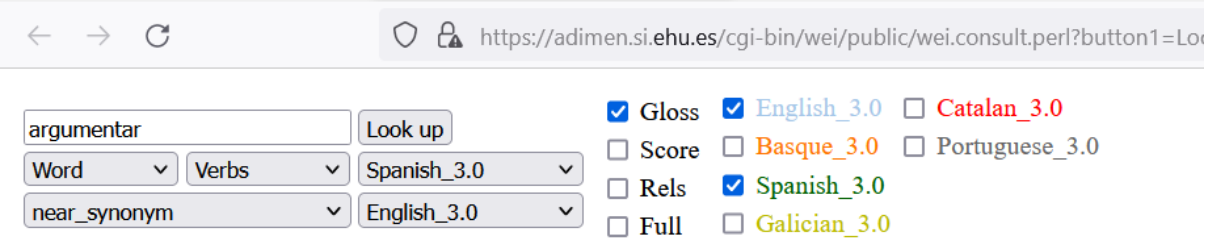

Figure 56. Search box in Multilingual Repository of the verb *argumentar*[27]

[26] https://adimen.si.ehu.es/cgi-bin/wei/public/wei.consult.perl
[27] Please check that the verb is in lower case; otherwise it may not return any results.

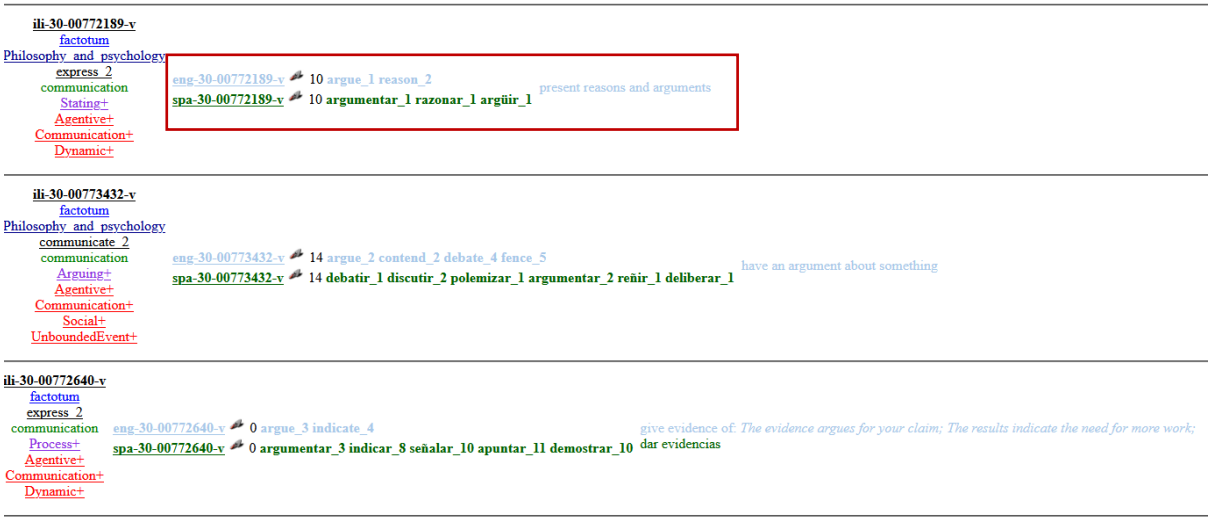

Figure 57. Senses in Multilingual Central Repository for *argumentar*

Please note that certain multiword expressions are also included in WordNet (e.g., *dar permiso*, *dar derecho a*). To search for them, use an underscore between words (i.e., *dar_permiso*, *dar_derecho_a*) (Figure 58):

Figure 58. Multiword expressions in WordNet

The task of the annotator is to go through the senses displayed and choose the one that better captures the sense of the particular CM. For example, for the meaning of *argumentar* in the class *argumentovat*/*argumentar*, sense 1 in WordNet (Figure 59):

Figure 59. Required information to add a link to Spanish WordNet

If no mapping is available, annotators should choose the option ‘NM’.

1.9 Examples

Each Spanish CM candidate is accompanied by a list of sentences extracted from the *XSRL* corpus, which are displayed in the *Examples* tab in the editor.

When selecting examples, it is important to select those that are illustrative of the sense of the Spanish verb. Please select 3-5 sentences (if possible) that better capture the meaning of the verb. Since the Spanish data extracted from the *XSRL* corpus consist of automatically translated texts, filtering sentences is crucial. Annotators should choose only grammatically correct sentences. Otherwise, they should discard sentences if these are too long or complicated, if they contain translation errors or if the content is inappropriate.

To add an example to the editor, the annotator should first check that the appropriate CM in the *Class Members* box and its example in the *Example* tab are selected and then click *Add to Lexicon* to assign the example sentence to the appropriate CM sense. Sentences that are assigned as examples are marked by an asterisk (*) (Figure 60).

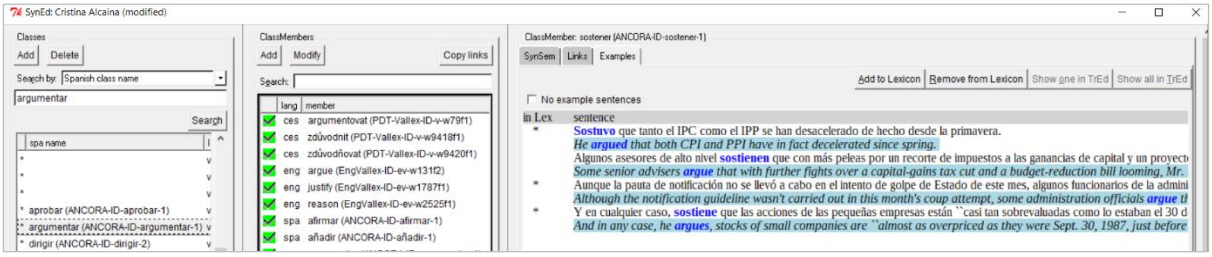

Figure 60. Assignment of the selected example sentences for the CM *sostener* (ANCORA-ID-sostener-1)

If no representative examples for the given CM are found, check the *No example sentences* box (Figure 61).

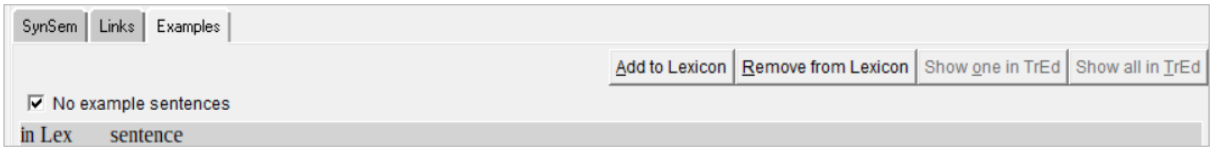

Figure 61. No example sentence box

1.10 Save your work

Once you finish working with SynEd, it is important that you save your progress. You can click ‘Save’ (bottom right corner in SynEd) or, alternatively, you can save your changes by clicking *Yes* in the message window that opens before exiting the program if you have made changes to the lexicon (Figure 62).

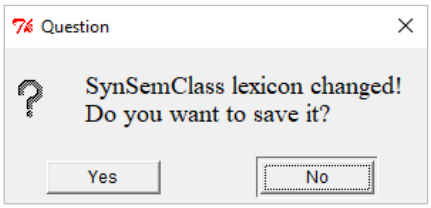

Figure 62. Message window to save progress

Apart from saving your work in SynEd and every time you make changes to the lexicon, it is important to commit those changes to TortoiseSVN. To do this, right-click on the folder *DataSynSemClass_multi* and click 'SVN Commit' (Figure 63). Please do not forget to write a brief description of the changes made in the *Message* box to make easier keeping a record of the annotation process.

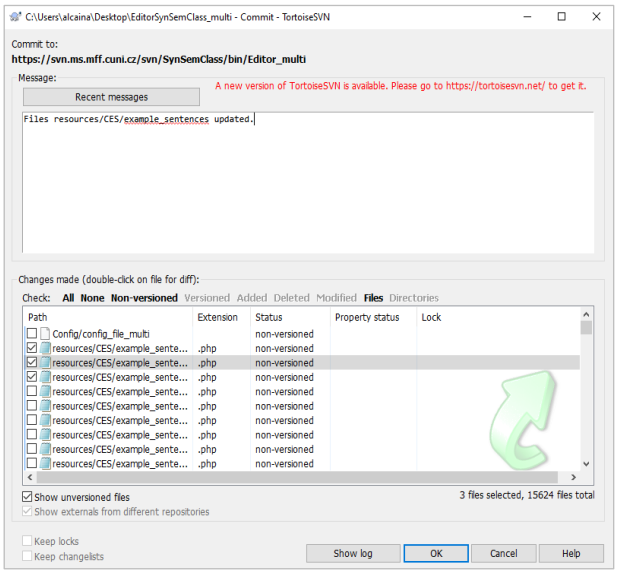

Figure 63. SVN Commit window

After committing the changes made, do not forget to update your folders. Open the menu by right-clicking the folder you want to update and click 'SVN Update'.

PART 2: TECHNICAL GUIDELINES

This part describes the procedure for the download and installation of the tools and packages required for the annotation of synonymous verbs in *SynSemClass*.

2.1 Download Setup Package

To download the setup package, go to https://ufal.mff.cuni.cz/tred, and download the file t*red-Installer-perl-included.exe* (from the Download section). This file contains tred, Strawberry Perl and required Modules.

Windows

Tred-installer-perl-included.exe (108M) - tred + Modules + Strawberry Perl

This is a full installation package of tred for Microsoft Windows. This installer contains tred, required Perl Modules and a recent version of Strawberry Perl.
Tred-installer.exe (30M) - tred + Modules

2.2 Installation

2.2.1 Installation of tred

The steps for the installation of tred in Windows are the following:

1. Go to the folder where tred has been downloaded and start the installation by running the downloaded file, named as *trend-installer-perl-included* (Figure 64).

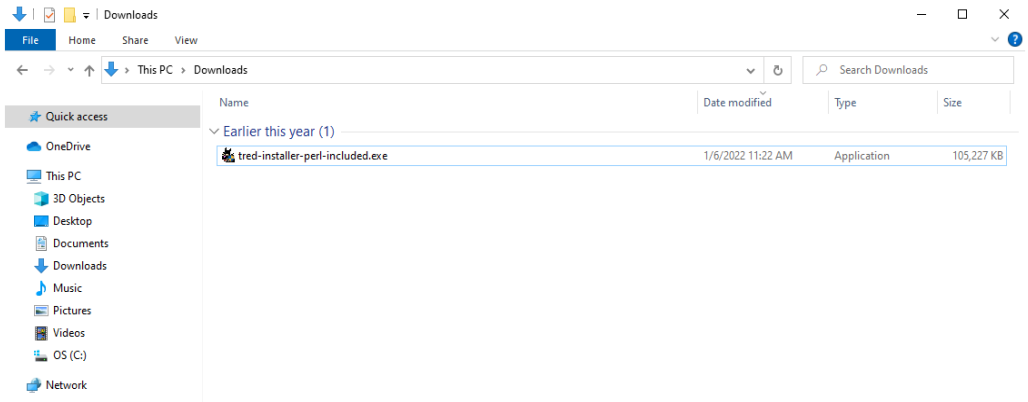

Figure 64. Tred installer

2. If Perl is not installed in your computer, the tred Installer will offer you to install it (you can confirm this by clicking the 'Install' button) (Figure 65).

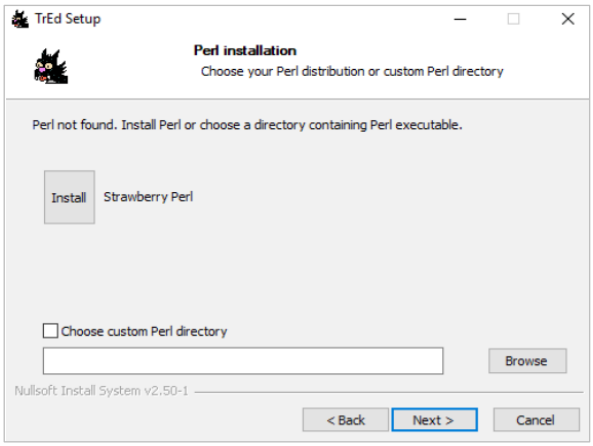

Figure 65. Perl installation module

3. After installing Perl (or if it was already installed in our system), the window displayed in Figure 66 opens:

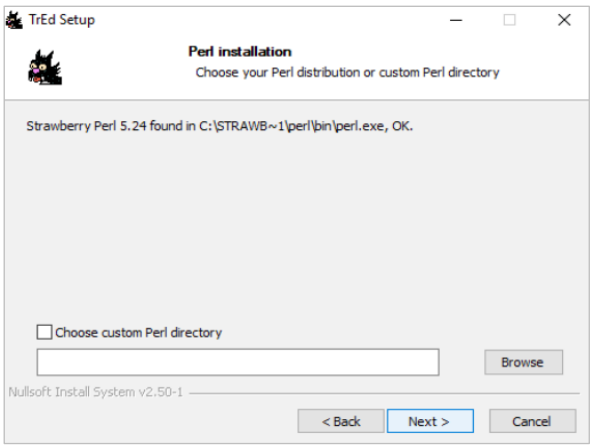

Figure 66. Perl installation module

4. Click ‘Next’ (Figure 63) to continue with the installation of tred. After installing tred, you must edit the *tred.bat* file in the *tred* directory. To do this, modify the last line by adding %TRED_DIR% before *tred* (i.e., ...\perl.exe %TRED_DIR%\tred %*) (in red in Figure 67).

Figure 67. Add %TRED_DIR% in tred

2.2.2 *Installation of additional Perl modules*

1. Run the *Command Prompt*. You can access it by using the search box in the Windows taskbar (Figure 68).

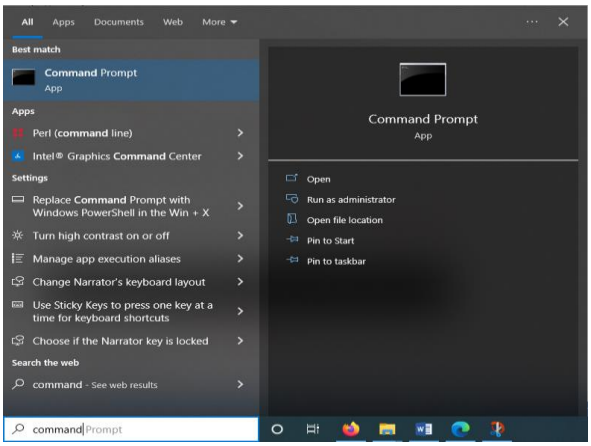

Figure 68. Command Prompt in Windows

2. Go to the *c:*\directory (command *cd c:*\) and then use the command ***cpan module_name*** (e.g., C:\>*cpan UNIVERSAL::DOES*) to install the required Perl modules (Figure 69).

Figure 69. *Cpan module_name in c:*\

List of required Perl modules:

- *UNIVERSAL::DOES*
- *Readonly*
- *Tk*
- *Treex::PML::Schema::CDATA*
- *XML::libxml::Iterator*
- *URI::Encode*
- *Tie::ixhash*

2.2.3 *Run tred and Install tred's Extensions*

You start to run tred by using the ***tred.bat file*** (as edited in point 4 in section 2.2.1) in the ***tred*** directory. If it does not work, follow the previous step and run the ***Command Prompt*** in Windows, go to the *tred* directory and run the ***tred.bat*** file.

Any error messages that appear during the startup should be sent to Eva Fučíková (***fucikova@ufal.mff.cuni.cz***).

If tred is run successfully, the next step is to add tred extensions (the so-called *modules*) to work with our data. To do this, open the ***tred Upgraded*** menu (Figure 70) and choose ***Manage Extensions***.

In case some problems appear, it is also possible to add tred extensions by choosing the ***Setup→manageextension*** in the tred ***menu*** to open the ***manageextensions*** window. Once the window is open, click ***getnewextensions***.

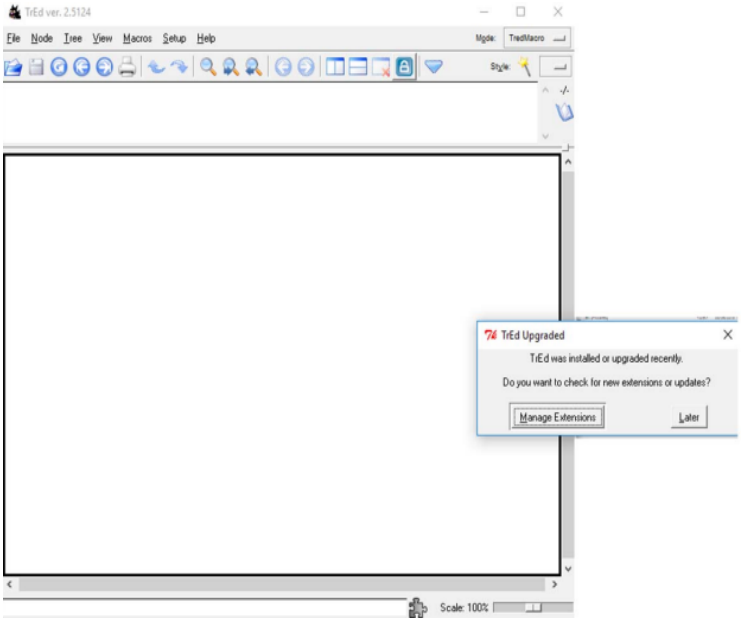

Figure 70. Manage extensions in tred

Both of the above-described procedures open the ***Install New Extensions*** window (Figure 68), where you can search for the necessary extensions by typing their names in the box ***Search***. The four extensions required are: ***pdt2.0***, ***pdt_vallex***, ***pedt***, and ***czengvallex***.

For these extensions, check the box ***Install*** (upper right corner in Figure 71). Once all the necessary extensions are selected, click '**Install Selected'** button (Figure 71).

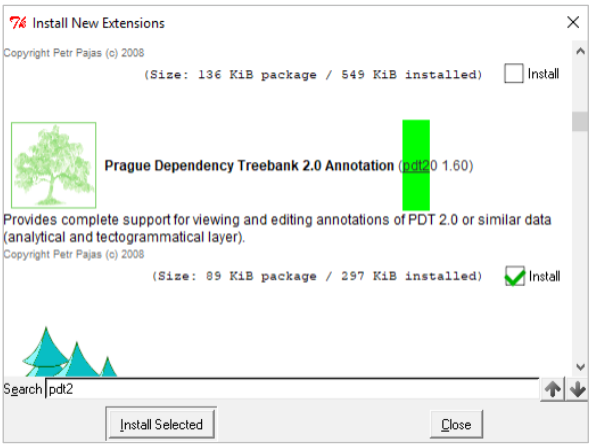

Figure 71. Install new extension window

2.3 Installation of Tortoise SVN

1. Go to https://tortoisesvn.net/downloads.html and download the TortoiseSVN version according to your system requirements (i.e., 32 bit or 64 bit) (Figure 72).

Figure 72. TortoiseSVN website

2. Go to the folder where TortoiseSVN has been downloaded and run the ***tortoisesvn….exe** file* (Figure 73):

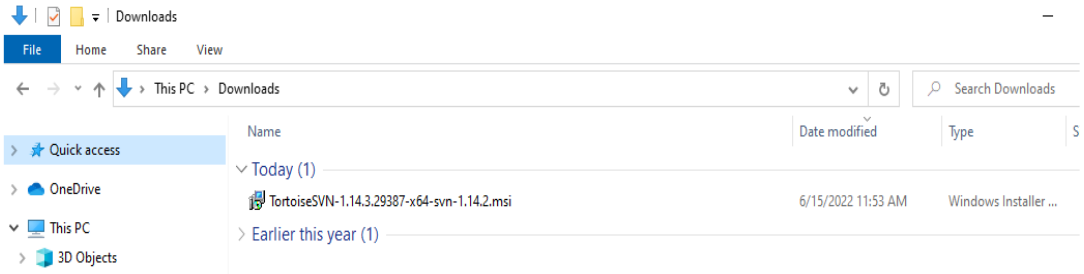

Figure 73. *TortoiseSVN.exe* file

2.4 Installation of SynEd

2.4.1 Editor

Create a new ***EditorSynSemClass_multi*** directory at the desktop (or, if preferred, at a different location in your system). Next, use right-click on this directory and select ***SVN Checkout*** from the menu. A ***Checkout*** window will open (Figure 74).

1. Fill in the ***URL of repository*** box with the address https://svn.ms.mff.cuni.cz/svn/SynSemClass/bin/Editor_multi.

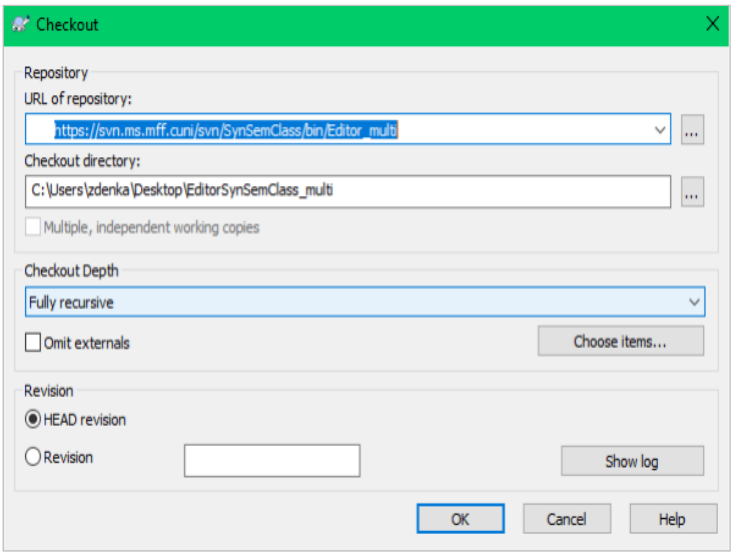

Figure 74. Checkout windows

When you press '**OK'**, you will be asked for credentials (they will be sent to you by email) (Figure 75).

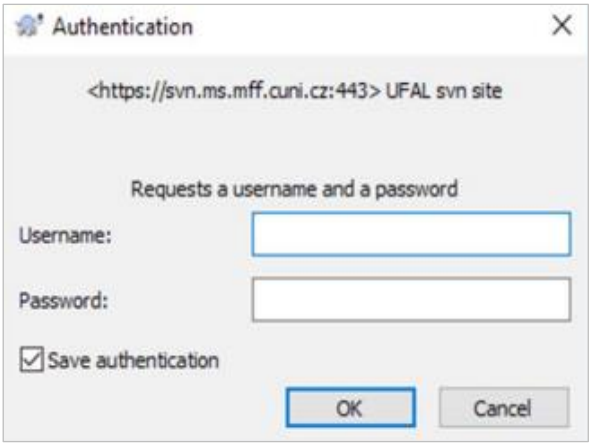

Figure 75. Authentication window

In the next step, go to the folder ***SynSemClasseditor_multi/resources***, where you can find the individual folders for each of the languages contained in the editor, at the moment, CES, DEU, ENG and SPA (Figure 76).

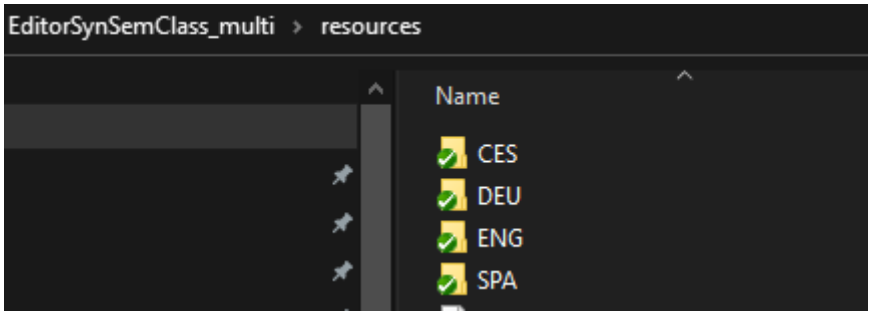

Figure 76. Languages available in the folder *EditorSynSemClass_multi*

Choose one language, e.g., ENG, and open the folder ***example_sentences*** (Figure 77).

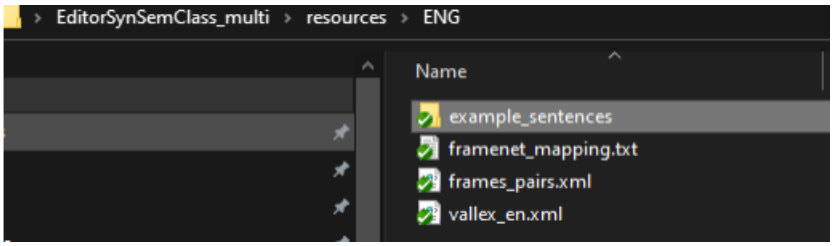

Figure 77. Open the *examples_sentences* folder

Extract the ***examples_lang.zip*** file, e.g., for English, the file is named as *examples_eng.zip* (Figure 78).

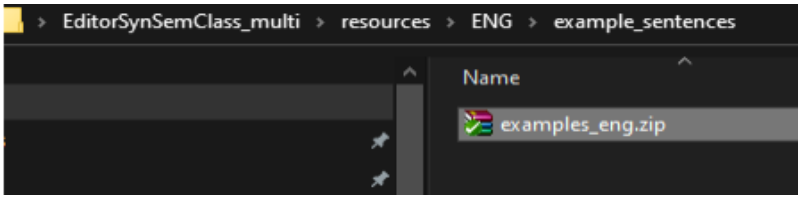

Figure 78. *Examples_eng.zip*

Right-click on the file *examples_lang.zip* to open the menu and choose **'Extract Here'** (it is important that the file is directly extracted in the *examples_sentences* folder and not in a new folder here) (Figure 79).

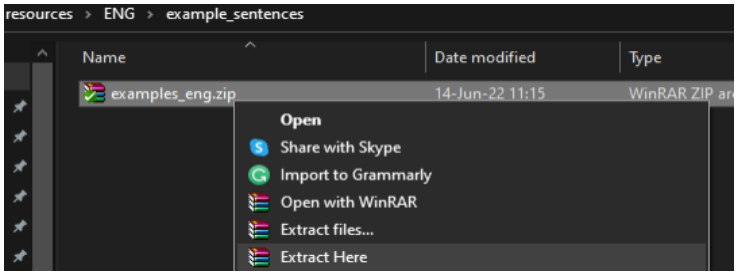

Figure 79. Examples extraction

Once the files are extracted, the folder should look similar to Figure 80:

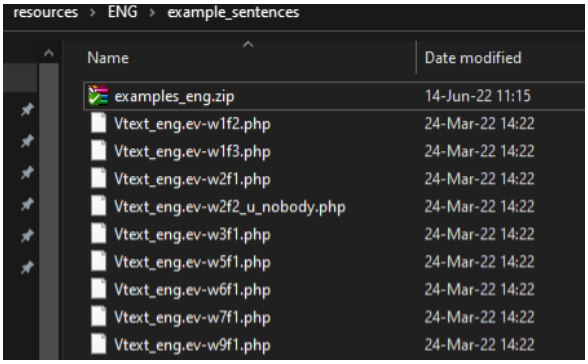

Figure 80. Example files extracted in the *examples_sentences* folder

2.4.2 Data

Each annotator needs to have its own annotation directory to store the processed data in. At the same place where you created the *EditorSynSemClass_mul*ti directory, create another directory called ***DataSynSemClass_multi*** and download a copy of our annotation directory in it. To do this, right-click on the folder, select ***SVN Checkout***, and fill in the following address in the ***URL of repository*** box in the Checkout window: https://svn.ms.mff.cuni.cz/svn/SynSemClass/data/anotace/AA. Please do not forget to replace 'AA' by your own initials (Figure 81).

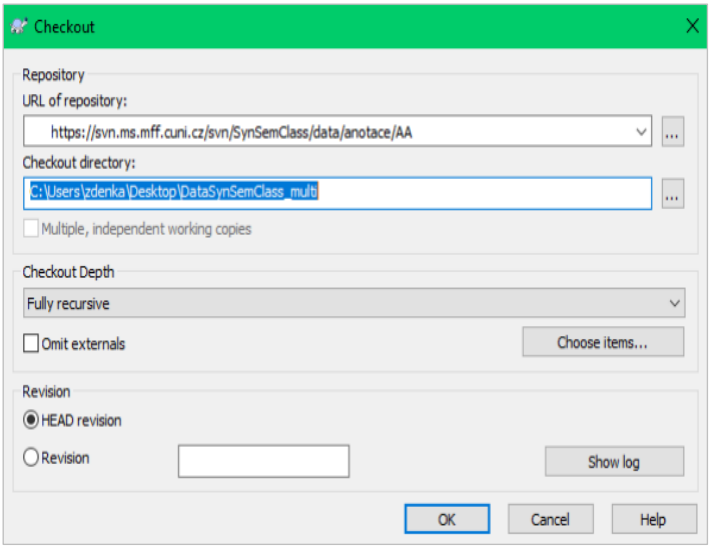

Figure 81. Creation of an annotator's directory

2.4.3 Configuration file settings

In the ***SynSemClasseditor_multi/Config*** directory, create a copy of the ***config_file_example*** and rename it as ***config_file_multi***) (Figures 82 to 84).

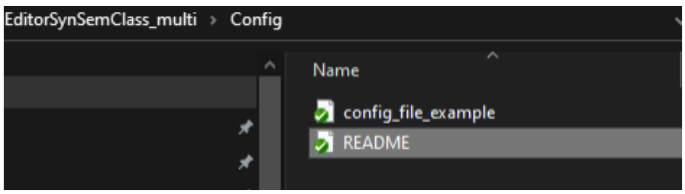

Figure 82. *SynsemClassEditor_multi/Config*

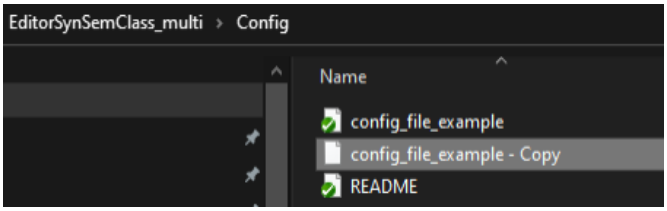

Figure 83. Create a copy of *config_file_example*

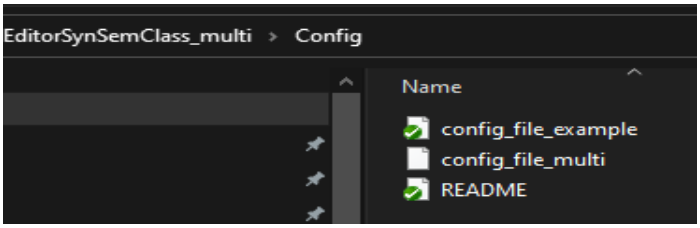

Figure 84. Rename the file as *config_file_multi*

The next step consists in setting the paths to the annotation file. To do this, right-click ***config_file_multi*** to open it in a text editor, e.g., Notepad (Figure 85).

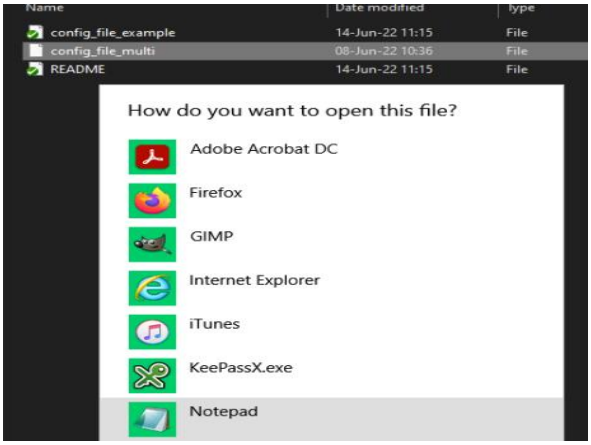

Figure 85. Open Notepad (or a similar text editor)

The proper path is as follows:

Resourcepath="c:\\Users\\Zdenka\\Desktop\\dataSynSemClass_multi,/c:\\Users\\Zdenka\\Desktop\\editorSynSemClass_multi\\resources"

Tredpath="C:\Tred32\tred\tred.bat"

Please, do remember to check that the path contains the actual path to the directory in your system (**in red** above).

See Figure 86 for the proper path settings for the following languages: CES, ENG, DEU, SPA[28].

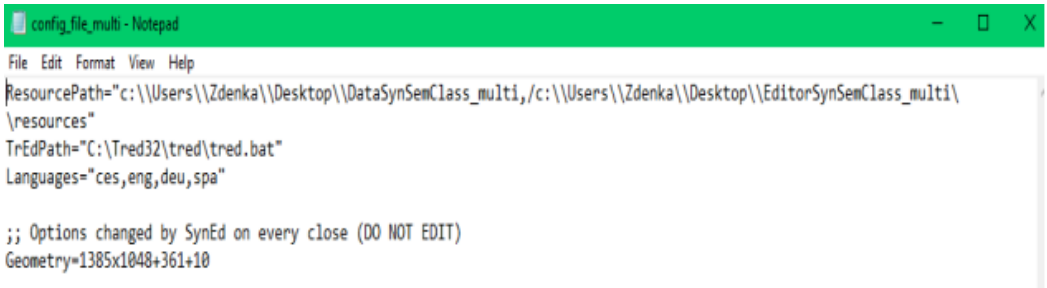

Figure 86. Path settings for languages CES, ENG, DEU and SPA

[28] Please note that the first language in the list is set as the default.

Choose the language you need to annotate (language codes are according to ISO norm Nr. 639, https://iso639-3.sil.org/code_tables/639/data). For now, there are the following languages available: CES, ENG, DEU, SPA. Be aware that the order of the languages determines the order of the synonym members in the editor and the language of the definitions. The choice and the order of languages is up to you. The language format is shown in Figure 86 (between quote marks and in lower case).

2.4.4 How to run the Multi version of SynSemClass Editor

To run the **Multi version of SynSemClass Editor** (*SynEd*), open the folder *SynSemClass_multi* located in the folder *EditorSynSemClass_multi* (Figure 87).

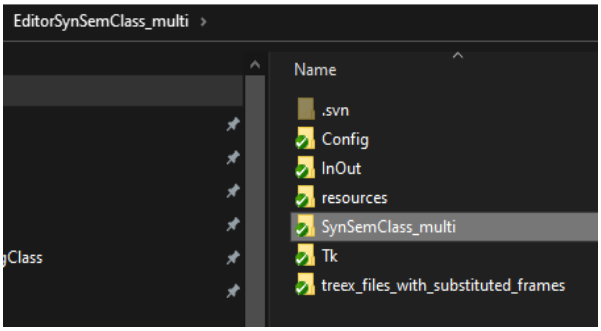

Figure 87. Open the folder *SynSemClass_multi*

In the folder *SynSemClass_multi*, double-click the file ***SynSemClass.bat*** (Figure 88):

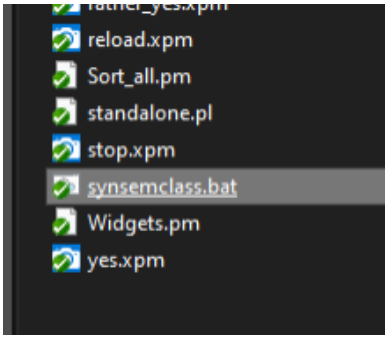

Figure 88. *Synsemclass.bat* file

It is important to follow the **correct** order of the paths listed. If the paths are set correctly, your name will appear on the top bar of the *SynSemClasseditor_multi*, (e.g., Zdenka Uresova) (Figure 89).

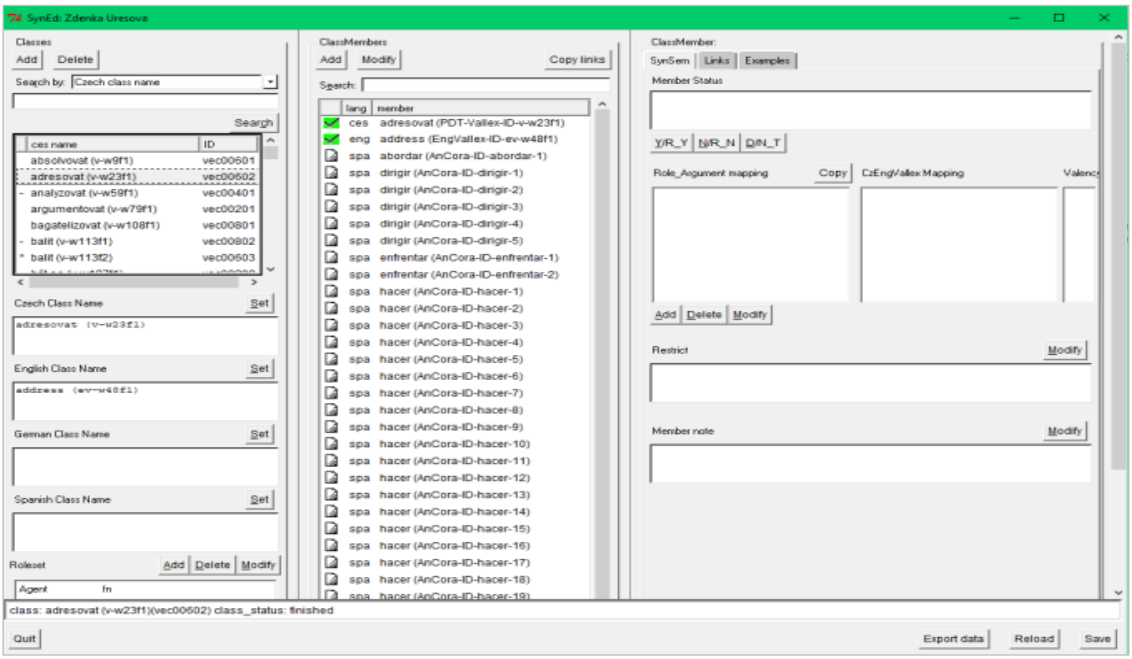

Figure 89. Syned

If problems come up, please repeat the steps described in section 2.4 (run the *Command Prompt*, change the directory to ***EditorSynSemClass_multi\SynSemClass*** and then run the *SynSemClass.bat* file contained in there) and send the error message that appears to Eva Fučíková.

APPENDIX

Lexical Semantic Structures in AnCora (adapted from Taulé et al. 2011: 5–12)

LSS A1: transitive-causative		**LSS A2: transitive-agentive**		**A3: ditransitive-agentive**			
A11. Transitive-causative	Arg0(cau) Arg1(tem)	A21. Transitive-agentive-patient	Arg0(agt) Arg1(pat)	A31. Distransitive-patient-locative	Arg0(agt) Arg1(pat) Arg2(loc)		
A12. Transitive-causative-state	Arg0(cau) Arg1(tem) Arg2(efi)	A22. Transitive-agentive-theme	Arg0(agt) Arg1(tem	A32. Ditransitive-patient-benefactive	Arg0(agt) Arg1(pat) Arg2(ben)		
A13. Transitive-causative-instrumental	Arg0(cau) Arg1(tem) Arg2(ins)	A23. Transitive-agentive-extension	Arg0(agt) Arg1(ext)	A33. Ditransitive-theme-locative	Arg0(agt) Arg1(tem) Arg2(loc)		
				A34. Ditransitive-patient-theme	Arg0(agt) Arg1(pat) Arg2(tem)		
				A35. Ditransitive-theme-cotheme	Arg0(agt) Arg1(tem) Arg2(cot)		
LSS B1: unacusative-motion		**LSS B2: unaccusative-state**					
B11. Unaccusative-motion	Arg1(tem) Arg2(loc)	B21. Unaccusative-state	Arg1(tem) Arg2(efi)				
B12. Unacusative-passive-ditransitive	Arg1(pat) Arg2(loc/ben/tem) Ar0(agt)	B22. Unaccusative-passive-transitive	Arg1(pat) Arg2(sgt)				
		B23. Unaccusative-cotheme	Arg1(tem) Arg2(cot)				
LSS C1: state-existential		**LSS C2: state-attributive**		**LSS C3: state-scalar**		**LSS C4: state-benefactive**	
C11. State-existential	Arg1(tem) Arg2(loc)	C21. State-attributive	Arg1(tem) Arg2(atr)	LSS C31. State-scalar	Arg1(tem) Arg(ext)	C41. State-benefactive	Arg1(tem) Arg2(ben)
						C42. State-experiencer	Arg1(tem) Arg2(exp)
LSS D1: inergative-agentive		**LSS D2: inergative-experiencer**		**LSS D3: inergative-source**			
D11. Inergative-agentive	Arg0(agt)	D21. Inergative-experiencier	Arg(exp)	D31. Inergative-source	Arg0 (src)		

The ÚFAL Technical Report Series

ÚFAL (Ústav formální a aplikované lingvistiky; https://ufal.mff.cuni.cz) is the Institute of Formal and Applied linguistics, at the Faculty of Mathematics and Physics of Charles University, Prague, Czech Republic. The Institute was established in 1990 after the political changes as a continuation of the research work and teaching carried out by the former Laboratory of Algebraic Linguistics since the early 60s at the Faculty of Philosophy and later the Faculty of Mathematics and Physics. Together with the "sister" Institute of Theoretical and Computational Linguistics (Faculty of Arts) we aim at the development of teaching programs and research in the domain of theoretical and computational linguistics at the respective Faculties, collaborating closely with other departments such as the Institute of the Czech National Corpus at the Faculty of Philosophy and the Department of Computer Science at the Faculty of Mathematics and Physics.

The ÚFAL technical report series has been established with the aim of disseminate topical results of research currently pursued by members, cooperators, or visitors of the Institute. Since November 1996, the following reports have been published.

ÚFAL TR-1996-01 Eva Hajičová, *The Past and Present of Computational Linguistics at Charles University*
Jan Hajič and Barbora Hladká, *Probabilistic and Rule-Based Tagging of an Inflective Language – A Comparison*

ÚFAL TR-1997-02 Vladislav Kuboň, Tomáš Holan and Martin Plátek, *A Grammar-Checker for Czech*

ÚFAL TR-1997-03 Alla Bémová at al., Anotace na analytické rovině, *Návod pro anotátory (in Czech)*

ÚFAL TR-1997-04 Jan Hajič and Barbora Hladká, *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structural Tagset*

ÚFAL TR-1998-05 Geert-Jan M. Kruijff, *Basic Dependency-Based Logical Grammar*

ÚFAL TR-1999-06 Vladislav Kuboň, *A Robust Parser for Czech*

ÚFAL TR-1999-07 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (in Czech)*

ÚFAL TR-2000-08 Tomáš Holan, Vladislav Kuboň, Karel Oliva, Martin Plátek, *On Complexity of Word Order*

ÚFAL/CKL TR-2000-09 Eva Hajičová, Jarmila Panevová and Petr Sgall, *A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-10 Zdeněk Žabokrtský, *Automatic Functor Assignment in the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-11 Markéta Straňáková, *Homonymie předložkových skupin v češtině a možnost jejich automatického zpracování*

ÚFAL/CKL TR-2001-12 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (III. verze)*

ÚFAL/CKL TR-2002-13 Pavel Pecina and Martin Holub, *Sémanticky signifikantní kolokace*

ÚFAL/CKL TR-2002-14 Jiří Hana, Hana Hanová, *Manual for Morphological Annotation*

ÚFAL/CKL TR-2002-15 Markéta Lopatková, Zdeněk Žabokrtský, Karolína Skwarská and Vendula Benešová, *Tektogramaticky anotovaný valenční slovník českých sloves*

ÚFAL/CKL TR-2002-16 Radu Gramatovici and Martin Plátek, *D-trivial Dependency Grammars with Global Word-Order Restrictions*

ÚFAL/CKL TR-2003-17 Pavel Květoň, *Language for Grammatical Rules*

ÚFAL/CKL TR-2003-18 Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska, Václava Benešová, *Valency Lexicon of Czech Verbs VALLEX 1.0*

ÚFAL/CKL TR-2003-19 Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, Oliver Čulo, *Anotování koreference v Pražském závislostním korpusu*

ÚFAL/CKL TR-2003-20 Kateřina Veselá, Jiří Havelka, *Anotování aktuálního členění věty v Pražském závislostním korpusu*

ÚFAL/CKL TR-2004-21 Silvie Cinková, *Manuál pro tektogramatickou anotaci angličtiny*

ÚFAL/CKL TR-2004-22 Daniel Zeman, *Neprojektivity v Pražském závislostním korpusu (PDT)*

ÚFAL/CKL TR-2004-23 Jan Hajič a kol., *Anotace na analytické rovině, návod pro anotátory*

ÚFAL/CKL TR-2004-24 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2004-25 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *The Prague Dependency Treebank, Annotation on tectogrammatical level*

ÚFAL/CKL TR-2005-27 Jiří Hana, Daniel Zeman, *Manual for Morphological Annotation (Revision for PDT 2.0)*

ÚFAL/CKL TR-2005-28 Marie Mikulová a kol., *Pražský závislostní korpus (The Prague Dependency Treebank) Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2005-29 Petr Pajas, Jan Štěpánek, *A Generic XML-Based Format for Structured Linguistic Annotation and Its application to the Prague Dependency Treebank 2.0*

ÚFAL/CKL TR-2006-30 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Annotation manual)*

ÚFAL/CKL TR-2006-31 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Anotace na tektogramatické rovině Pražského závislostního korpusu (Referenční příručka)*

ÚFAL/CKL TR-2006-32 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall,Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Reference book)*

ÚFAL/CKL TR-2006-33 Jan Hajič, Marie Mikulová, Martina Otradovcová, Petr Pajas, Petr Podveský, Zdeňka Urešová, *Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené řeči*

ÚFAL/CKL TR-2006-34 Markéta Lopatková, Zdeněk Žabokrtský, Václava Benešová (in cooperation with Karolína Skwarska, Klára Hrstková, Michaela Nová, Eduard Bejček, Miroslav Tichý) *Valency Lexicon of Czech Verbs. VALLEX 2.0*

ÚFAL/CKL TR-2006-35 Silvie Cinková, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, Zdeněk Žabokrtský, *Annotation of English on the tectogrammatical level*

ÚFAL/CKL TR-2007-36 Magda Ševčíková, Zdeněk Žabokrtský, Oldřich Krůza, *Zpracování pojmenovaných entit v českých textech*

ÚFAL/CKL TR-2008-37 Silvie Cinková, Marie Mikulová, *Spontaneous speech reconstruction for the syntactic and semantic analysis of the NAP corpus*

ÚFAL/CKL TR-2008-38 Marie Mikulová, *Rekonstrukce standardizovaného textu z mluvené řeči v Pražském závislostním korpusu mluvené češtiny. Manuál pro anotátory*

ÚFAL/CKL TR-2008-39 Zdeněk Žabokrtský, Ondřej Bojar, *TectoMT, Developer's Guide*

ÚFAL/CKL TR-2008-40 Lucie Mladová, *Diskurzní vztahy v češtině a jejich zachycení v Pražském závislostním korpusu 2.0*

ÚFAL/CKL TR-2009-41 Marie Mikulová, *Pokyny k překladu určené překladatelům, revizorům a korektorům textů*

z Wall Street Journal pro projekt PCEDT

ÚFAL/CKL TR-2011-42 Loganathan Ramasamy, Zdeněk Žabokrtský, *Tamil Dependency Treebank (TamilTB) – 0.1 Annotation Manual*

ÚFAL/CKL TR-2011-43 Ngụy Giang Linh, Michal Novák, Anna Nedoluzhko, *Coreference Resolution in the Prague Dependency Treebank*

ÚFAL/CKL TR-2011-44 Anna Nedoluzhko, Jiří Mírovský, *Annotating Extended Textual Coreference and*

Bridging Relations in the Prague Dependency Treebank

ÚFAL/CKL TR-2011-45 David Mareček, Zdeněk Žabokrtský, *Unsupervised Dependency Parsing*

ÚFAL/CKL TR-2011-46 Martin Majliš, Zdeněk Žabokrtský, *W2C – Large Multilingual Corpus*

ÚFAL TR-2012-47 Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Zuzanna Bedřichová, Jiří Mírovský,

Magdaléna Rysová, Jana Zdeňková, Veronika Pavlíková, Eva Hajičová,

Manual for annotation of discourse relations in the Prague Dependency Treebank

ÚFAL TR-2012-48 Nathan Green, Zdeněk Žabokrtský, *Ensemble Parsing and its Effect on Machine Translation*

ÚFAL TR-2013-49 David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Daniel Zemana, Zdeněk Žabokrtský, Jan Hajič *Cross-language Study on Influence of Coordination Style on Dependency Parsing Performance*

ÚFAL TR-2013-50 Jan Berka, Ondřej Bojar, Mark Fishel, Maja Popović, Daniel Zeman, *Tools for Machine Translation Quality Inspection*

ÚFAL TR-2013-51 Marie Mikulová, *Anotace na tektogramatické rovině. Dodatky k anotátorské příručce (s ohledem na anotování PDTSC a PCEDT)*

ÚFAL TR-2013-52 Marie Mikulová, *Annotation on the tectogrammatical level. Additions to annotation manual (with respect to PDTSC and PCEDT)*

ÚFAL TR-2013-53 Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský,
Úpravy a doplňky Pražského závislostního korpusu (Od PDT 2.0 k PDT 3.0)

ÚFAL TR-2013-54 Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský,
From PDT 2.0 to PDT 3.0 (Modifications and Complements)

ÚFAL TR-2014-55 Rudolf Rosa, *Depfix Manual*

ÚFAL TR-2014-56 Veronika Kolářová, *Valence vybraných typů deverbativních substantiv ve valenčním slovníku PDT-Vallex*

ÚFAL TR-2014-57 Anna Nedoluzhko, Eva Fučíková, Jiří Mírovský, Jiří Pergler, Lenka Šíková,
Annotation of coreference in Prague Czech-English Dependency Treebank

ÚFAL TR-2015-58 Zdeňka Urešová, Eva Fučíková, Jana Šindlerová,
CzEngVallex: Mapping Valency between Languages

ÚFAL TR-2015-59 Kateřina Rysová, Magdaléna Rysová, Eva Hajičová,
Topic–Focus Articulation in English Texts on the Basis of Functional Generative Description

ÚFAL TR-2016-60 Kira Droganova, Daniel Zeman,
Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies

ÚFAL TR-2018-61 Lukáš Kyjánek,
Morphological Resources of Derivational Word-Formation Relations

ÚFAL TR-2019-62 Zdeňka Urešová, Eva Fučíková, Eva Hajičová,
CzEngClass: Contextually-based Synonymy and Valency of Verbs in a Bilingual Setting (CzEngClass: Kontextová synonymie a valence sloves v bilingvním prostředí)

ÚFAL TR-2019-63 Ján Faryad,
Identifikace derivačních vztahů ve španělštině

ÚFAL TR-2020-64 Marie Mikulová, Jan Hajič, Jíří Hana, Hana Hanová, Jaroslava Hlaváčová, Emil Jeřábek, Barbora Štěpánková, Barbora Vidová Hladká, Daniel Zeman,
Manual for Morphological Annotation. Revision for Prague Dependency Treebank – Consolidated 2020 release

ÚFAL TR-2021-65 Rudolf Rosa, *Technická zpráva o vývoji projektu THEaiTRE v roce 2020*

ÚFAL TR-2021-66 Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Daniel Zeman
Coreference meets Universal Dependencies – a pilot experiment on harmonizing coreference datasets for 11 languages

ÚFAL TR-2021-67 THEaiTRobot 1.0, David Košťák, Daniel Hrbek, Rudolf Rosa, Ondřej Dušek,
AI: When a Robot Writes a Play

ÚFAL TR-2021-68 Markéta Lopatková, Václava Kettnerová, Anna Vernerová, Eduard Bejček, Zdeněk Žabokrtský,
Valenční slovník českých sloves VALLEX

ÚFAL TR-2021-69 Niyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda, Madga Ševčíková, Jonáš Vidra, Zdeněk Žabokrtský
Towards Universal Segmentations: Survey of Existing Morphosegmentation Resources

ÚFAL TR-2021-70 Zdeňka Urešová, Eva Fučíková, Jan Hajič, Karolina Zaczynska,
Annotation guidelines for German verbal synonyms included in SynSemClass Lexicon

ÚFAL TR-2022-71 THEaiTRobot 2.0, Josef Doležal, Klára Vosecká, Tomáš Musil, David Mareček, Rudolf Rosa,
Permeation / Prostoupení

ÚFAL TR-2022-72 Cristina Fernández-Alcaina, Eva Fučíková, Zdeňka Urešová,
Annotation guidelines for Spanish verbal synonyms in the SynSemClass Lexicon