FROM CONCEPT DEFINITIONS TO SEMANTIC ROLE LABELING IN SPECIALIZED KNOWLEDGE RESOURCES

Ivana Brač and Ana Ostroški Anić

Institute of Croatian Language and Linguistics

Abstract

The paper presents the framework for semantic role labeling that is being developed within the project the *Dynamicity of Specialized Knowledge Categories*. The main goal of the project is the description of the categories of aviation within semantic frames. One of the tasks towards achieving this goal is developing a methodology for the syntactic and semantic analysis of the language for specialized purposes that could be applied to any specialized domain. The paper first outlines the differences between the terminological database Struna and a lexical database of semantic frames of aviation, AirFrame. Methods applied in analyzing terminological definitions and annotated sentences containing aviation terms are then presented. Conceptual information from the definitions of key aviation concepts in Struna is compared to semantically annotated sentences extracted from the parallel English-Croatian aviation corpus. The corpus compilation and analysis has been done in Sketch Engine, while the annotation is done in the WebAnno tool. Two approaches to semantic role labelling are discussed. The first approach does not have roles specific to one class, e.g. LIRICS (Petukhova & Bunt, 2008) and VerbNet (Kipper et al., 2008) with hierarchically organized semantic roles. The second approach, developed within FrameNet has more than 2000 frame elements or semantic roles that are verb-specific and frame-specific. Advantages of each approach are discussed, and the semantic tagset applied in the analysis is elaborated. The benefits of using semantic role labeling in terminological resources are finally discussed.

Keywords: terminology, semantic roles, semantic role labeling, specialized knowledge.

1. Introduction[76]

The development of specialized knowledge resources relies heavily on using the existing tools for general language description, and on integrating good lexicographic practice and theoretical findings in such a way so as to establish a link between general and specialized knowledge. Online terminological dictionaries and databases that exploit the semantic-syntactic potential of terminological units, thus trying

[76] This work has been fully supported by the Croatian Science Foundation (www.hrzz.hr) under the project UIP-2017-05-7169.

to reflect the nature of specialized knowledge categories, are becoming more and more recognized (Faber, 2012; L'Homme, 2012,). Often is the work of developing a specialized database of a particular domain the continuation of previous, more traditionally oriented terminology work, as is the case of AirFrame, a database of semantic frames in the field of aviation that is being created within the research project the *Dynamicity of Specialized Knowledge Categories* (DIKA). One of the tasks towards achieving this goal is developing a methodology for the syntactic and semantic analysis of the language for specialized purposes that could be applied to any specialized domain. In order to compile a set of semantic roles applied in the process, a comparison between terminological definitions of aviation concepts in Struna and sentences extracted from a specialized corpus has been made.

Struna is a terminological database developed within the program the *Development of the Croatian Special Field Terminology* (known under its Croatian acronym Struna), financed by the Croatian Science Foundation (HRZZ). The program started in 2007, and is being carried out at the Institute of Croatian Language and Linguistics. The main purpose of Struna is to standardize Croatian terminology of various professional domains, and make it available to the public through a national term bank developed in-house for this purpose (Brač, Bratanić, & Ostroški Anić, 2015). As most normative term banks the primary task of which is defining and prescribing terminology in languages with a less standardized terminological component, Struna is largely based on the traditional terminological premises set out in the work of Eugen Wüster and his followers. According to these terminological principles, codified in the ISO terminology standards, the terminology of a certain special field is defined as a structured group of concepts, concept relations and terms as their designations (ISO 704). However, although the model of the terminology standardization applied in Struna originally rather strictly adhered to the semiotic principles defined by this terminological tradition, certain accommodations have been made over the course of years, both in data categories and the methodology applied, that moved Struna more in the direction of a descriptive end of terminology work.

The organization of the terminology workflow required a proper terminology management system; therefore an in-house solution was designed in order to cover both the need for an editing and storage tool, as well as for a search and retrieval application. In order to make Struna compatible and exchangeable with the existing termbases, its structure was designed in accordance with the TEI P5 guidelines for text mark-up and the TBX standard format for the representation and exchange of terminological data (Melby, 2015). The current list of record elements includes the following data categories: subject field and subfield, preferred Croatian term, source of the term, foreign term language label, neologism label, interdisciplinary term label, grammatical information on the preferred term, definition and its source, context with its source, synonyms according to their normative status (admitted, proposed, deprecated, obsolete, colloquial), equivalents in other languages, subordinate concept, abbreviation in Croatian and other languages, symbol, formula, equation, hyperlink, picture, note, and a field for correspondence among the domain editors, terminologists, and language experts (Bratanić & Ostroški Anić, 2013).

Unlike Struna, which is a multidomain terminological database, AirFrame is a monodomain terminological resource. It is a lexical database with a terminological function in which aviation concepts and their definitions, terms in several languages, and other relevant categories of terminological entries are presented in specialized semantic frames. Such a presentation of specialized knowledge is dynamic because

it is based on situationally, contextually and culturally conditioned semantic frames (Fillmore, 1982; Fillmore, Johnson, & Petruck, 2003). The database will consist of these categories: semantic frames, frame elements with their definitions, terms and terminological units such as collocations and phraseological units (corresponding to lexical units in FrameNet), sentences labeled with semantic roles, and frame-to-frame conceptual relations. Following the general methodology of FrameNet (Ruppenhofer et al., 2010), semantic frames and their elements are defined separately as opposed to lexical data.

2. Methods

Definitions of key aviation concepts as defined in Struna were analyzed and compared to sentences containing terms for those concepts, in order to establish the difference in conceptual and semantic information that can be gathered from terminological definitions as opposed to annotated examples from corpora. The list of key aviation concepts used as target words for extracting sentences included: *aerodrome*, *aircraft*, *airline*, *airspace*, *air traffic*, *air traffic control*, *air transport*, *flight*, *landing* and *taking off*. Definitions were analyzed for concept characteristics, and annotated for ontological categories contained in them, which make frame elements. A parallel English-Croatian aviation corpus compiled within the project was then queried for English and Croatian terms of the analyzed aviation concepts. Extracted sentences were annotated for semantic roles using the LIRICS semantic tagset (Petukhova & Bunt, 2008). Corpus analysis was done in Sketch Engine (Kilgariff et al., 2014), while the annotation was carried out using the WebAnno tool (Eckart de Castilho et al., 2016).

3. Results

Since traditional terminology work is largely focused on defining entities as prototypical categories for knowledge representation, it was expected that entities would comprise the largest share of the types of categories labeled as frame elements in definitions. Apart from entities such as *aircraft*, *airplane*, *person* or *thing*, bounded regions like *place*, *area* and *aerodrome areas* also fall within a large group of entities, but are labeled as locations according to the terminology of FrameNet. Activities like *flying*, *take-off*, *movement* and *transport* are typically present in the field of aviation, as well as a number of procedures that we ontologically define as a type of an activity. Examples (1) to (5) show the definitions of five aviation concepts defined within the semantic frames Aerodrome, Air_transport and Flight, in which frame elements are marked in Italics. Figure 1. shows how sentences were annotated in WebAnno.

> (1) aerodrome – a defined *area* on land or water, including any *objects*, *installations and equipment*, intended to be used either wholly or in part for the movement, take-off, landing and parking of *aircraft*.

(2) airline – an air *operator* that uses *aircraft* to *transport* people and/or goods for *commercial purposes*.

(3) air transport – the *transport* of *people* or *things* from one place to another by means of an *aircraft*.

(4) flight – the *flying* of *aircraft* from any *aerodrome* to a *destination* aerodrome.

(5) take-off – an aircraft *operation* during which an *aircraft* accelerates from the stop phase, leaves the ground and reaches the required the *flight level*.

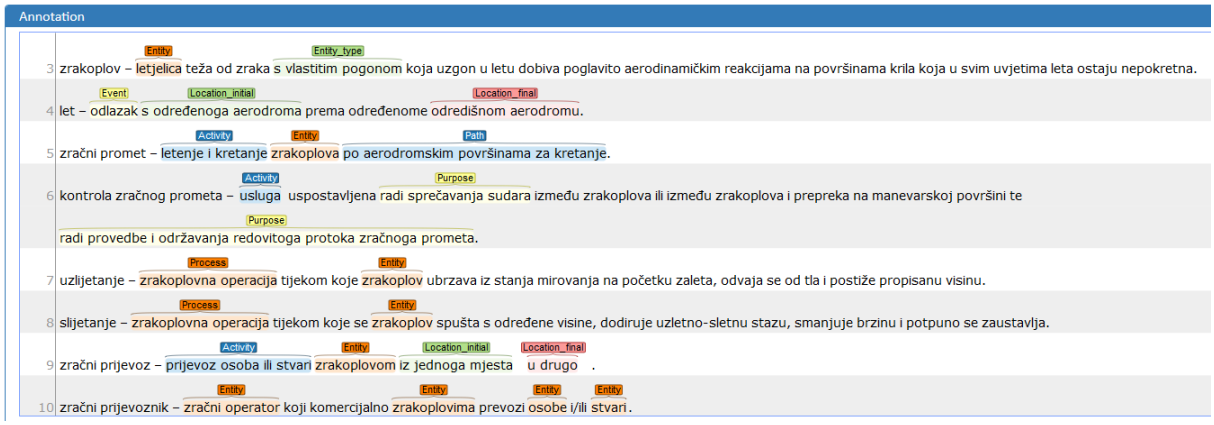

Figure 1. Annotated frame elements as represented in Croatian definitions extracted from Struna

The parallel English-Croatian aviation corpus compiled within the project was used to extract sentences used for annotation. The corpus is compiled from the Directory of legal acts of the European Union, from the chapter "Transport policy", subchapter Air transport in English and Croatian. Out of 220 documents from the "Air transport" subchapter, 178 legal acts are taken having both language versions. The texts are downloaded from the EUR-Lex database, and entered into the Sketch Engine's corpus compilation module. The corpus was queried for the target terms *aerodrome*, *aircraft*, *airline*, *airspace*, *air traffic*, *air traffic control*, *air transport*, *flight*, *landing* and *take-off*, and sentences were manually extracted and annotated. Examples (6) to (9) show some of the sentences with the target terms *aerodrome*, *aircraft*, *airline* and *flight*. Semantic roles are marked in small caps in square brackets following the sentence element they label.

(6) The flight previously notified by a basic flight data process [THEME] will now not enter the airspace of the notified unit [FINAL_LOC].

(7) An applicant [AGENT] shall fly the aircraft [THEME] from a position where the PIC functions can be performed [INITIAL_LOC] and to carry out the test [THEME] as if there is no other crew member [SETTING].

(8) An aircraft taxiing on the manoeuvring area of an aerodrome [AGENT] shall give way to aircraft taking off or about to take off [BENEFICIARY].

(9) Similarly, the airline operating the aircraft [PIVOT] needs underlying economic authority [THEME] from the DOT [SOURCE].

The terminology of the semantic roles in the LIRICS tagset obviously differs from FrameNet's roles in some aspects, but differences could be noticed between certain roles in other projects, e.g. VerbNet and PropBank. E.g., Initial location and Final location as used in LIRICS and in our annotation correspond to Source and Destination in VerbNet, while the Setting corresponds to the meaning labeled as Circumstances in FrameNet.

4. Discussion

Since semantic relations are crucial in connecting frame elements, and consecutively in determining the relations between different frames, they are reflected at the sentence level as verbs, their arguments and adjuncts, to which semantic roles are assigned. The number of semantic roles and their definitions vary depending on the degree of abstractedness and concreteness. Roughly, there are two approaches to semantic role labeling. The first approach does not have roles specific to one class, e.g. LIRICS (Petukhova & Bunt, 2008) and VerbNet (Kipper et al., 2008) with hierarchically organized semantic roles. The second approach, developed within FrameNet has more than 2000 frame elements or semantic roles that are verb-specific and frame-specific. Gantar et al. (2018), following and simplifying the Prague Dependency Treebank (PDT), use 25 semantic roles for the annotation of examples extracted from the general language corpora in Slovenian and Croatian. An even more reduced tagset (17 semantic roles) is applied in the creation of the semantic layer of the Croatian Dependency Treebank (Farkaš, Filko, & Tadić, 2016). Semantic role labeling in specialized knowledge resources differs to some extent, depending largely on the nature of conceptual relations within a particular domain. E.g., the semantic roles set used in Ecolexicon (Araúz et al., 2012) is based on the organization of the field of environment around *event* as the key category of the domain as well as the roles applied in the framed version of the Canadian database DiCoEnviro (L'Homme, Robichaud, & Rüggeberg 2014).

An ideal terminological intensional definition should contain a concept superordinate to the one that is being defined, and the delimiting characteristics, i.e. the characteristics that set out the defined concept from concepts similar or related to it. Ontological categories labeled in the analyzed definitions refer to frame elements that invoke particular semantic frames that structure the field of aviation. If we take the frame of Air_Transport as an example, it is defined as 'the transport of people or things from one place to another by means of an aircraft'. Transport, People, Things and Aircraft can be said to be the frame elements invoking this frame, and each one of them is then defined and put into relation with other semantic

frames they might also be part of. However, the relations among different frame elements in a frame are only implicitly present in the frame definition. A better insight into the complexity of both intraframe and interframe relations is gained at the lexical level, where frame elements are expressed as semantic roles.

Choosing a model and a specific tagset to be applied in semantic role labeling depends largely on the theoretical approach referred to. Although FrameNet offers an exhaustive list of roles, its magnitude also presents its largest drawback, i.e. a difficulty to correlate with resources that apply a more general and limited SRL tagset. Using less semantic roles in annotation, on the other hand, can lead to leaving out useful semantic information. Duration is thus a semantic role missing in the LIRICS tagset, but it is an important role for labeling examples from an aviation corpus because it marks a relevant component in activities like aircraft procedures. However, a more thorough analysis still needs to be carried out in order to reach a conclusion on the benefits and pitfalls of adding more specific semantic roles, as well as on the methodology applied.

5. Conclusion

Online terminological resources that refer to the organization and presentation of specialized knowledge according to dynamic categories like semantic frames mostly refer to FrameNet's methodology in defining semantic frames and frame elements. Some of them, like the Ecolexicon database, set up on the principles of Frame-based terminology (Faber et al. 2012), and the Canadian databases DiCoEnviro and DiCoInfo (L'Homme, Robichaud, & Rüggeberg 2014) have modified and simplified the FrameNet's semantic roles. Although the merging of resources and the interchange of data is a pressing need in contemporary digital lexicography, the purpose of each resource and the needs of its users must bear more relevance in deciding what methodology to apply. The LIRICS semantic tagset that was applied in the annotation of sentences from aviation related texts has to be enriched with a few semantic roles that are not present in the current list, e.g. Duration and Inanimate agent. However, the tagset should not be too large, otherwise it would slow down the process of annotation, without contributing much to the project goals. Categories of specialized knowledge can be defined in terms of semantic relations that bind them together, but it must nevertheless be done in a precise, clear and concise manner so to reflect both terminological needs and the need for good terminology.

References

Araúz, P. L. et al. 2012. Specialized language semantics. In P. Faber (Ed.), *A Cognitive Linguistics View of Terminology and Specialized Language* (pp. 95-176). Berlin/New York: De Gruyter Mouton.

Brač, I., M. Bratanić, & A. Ostroški Anić. (2015). Hrvatsko nazivlje i nazivoslovlje od Šuleka do Strune - hrvatski jezik i terminološko planiranje. In M. Bratanić, I. Brač, & B. Pritchard (Eds.), *Od Šuleka do Schengena: terminološki, terminografski i prijevodni aspekti jezika struke* (pp. 3-26). Zagreb/Rijeka: Institut za hrvatski jezik i jezikoslovlje & Pomorski fakultet u Rijeci.

Bratanić, M., & A. Ostroški Anić. (2013). The Croatian National Termbank STRUNA: A New Platform for Terminological Work. *Collegium Antropologicum, 37*(3), 677-683.

Eckart de Castilho, R. et al. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), Osaka, Japan* (pp. 76-84). Retrieved from https://pdfs.semanticscholar.org/aee7/1bfa28ec1bad78d4bd4aadcab168aa6b3b13.pdf

Faber, P. (Ed.) (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/Boston: De Gruyter Mouton.

Farkaš, D., M. Filko, & M. Tadić. (2016). HR4EU – Using Language Resources in Computer Aided Language Learning. *CLIB 2016 Proceedings* (pp. 38-46). Retrieved from https://bib.irb.hr/datoteka/852491.clib_farkas.pdf

Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the Morning Calm* (pp. 111-137). Seoul, South Korea: Hanshin Publishing Company.

Fillmore, C. J., Christopher R. J., & M. R. L. Petruck. (2003). Background to FrameNet. *International Journal of Lexicography, 16*(3), 235-250.

Gantar, P. et al. (2018). Towards Semantic Role Labeling in Slovene and Croatian. *Konferenca Jezikovne tehnologije in digitalna humanistika, Ljubljana*. Retrieved from http://www.sdjt.si/wp/wp-content/uploads/2018/09/JTDH-2018_Gantar-et-al_Towards-Semantic-Role-Labeling-in-Slovene-and-Croatian.pdf

ISO 704. (2000). *Terminology work – Principles and methods*. International Organization for Standardization.

Kilgarriff, A. et al. (2014). The Sketch Engine: ten years on. *Lexicography, 1*(1), 7-36.

Kipper, K. et al. (2008). A large-scale classification of English verbs. *Lang Resources & Evaluation, 42*, 21-40.

L’Homme, M. C. (2012). Adding syntactico-semantic information to specialized dictionaries: an application of the FrameNet methodology. *Lexicographica*, *23*, 233–252.

L’Homme, M. C, B. Robichaud, & C. S. Rüggeberg. (2014). Discovering Frames in Specialized Domains. In N. Calzolari et. Al (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 1364-1371). Reykjavik, Iceland: European Language Resources Association (ELRA).

Melby, A. K. (2015). TBX: A terminology exchange format for the translation and localization industry. In H. J. Kockaert & F. Steurs (Eds). *Handbook of Terminology, Volume 1* (p. 392-423). Amsterdam/Philadelphia: John Benjamins.

Petukhova, V., & H. Bunt. (2008). LIRICS semantic role annotation: design and evaluation of a set of data categories. *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (pp. 39-45). Retrieved from https://pdfs.semanticscholar.org/732c/65885e1e664c44db6ad723425547633dad7a.pdf

Ruppenhofer, J. et al. (2010). FrameNet II: Extended Theory and Practice. Retrieved from https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf