A parallel English-Croatian corpus in the domain of air traffic

Maja Lončar[1] and Ana Ostroški Anić[1]

[1]Institute of Croatian Language and Linguistics

INTRODUCTION

Dynamicity of Specialized Knowledge Categories (DIKA) is an installation research project, financed by the Croatian Science Foundation, dealing with the description of conceptual and linguistic levels of specialized knowledge categories within the dynamic nature of semantic frames. A parallel corpus of English and Croatian texts in the domain of air traffic has been compiled for the purposes of term and definition extraction as well as for the semantic and syntactic analysis of key terms in the domain of aviation.

Since the project's key output is a multilingual terminological database in which the aviation terminology is defined in semantic frames, including frame elements, conceptual relations, and figurative and phraseological terminological units, the parallel English-Croatian corpus of aviation terminology serves as the starting point for term extraction and term analysis. The subdomain of air traffic has been chosen as the most representative aspect of the vast and interdisciplinary field of aviation. All term extraction and term analysis has been done in Sketch Engine tools.

METHODOLOGY

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 "Air transport" subchapter, 178 legal acts are taken having both (English and Croatian) language versions. The texts are downloaded from the EUR-Lex database[1] and entered into the Sketch Engine's corpus compilation module [2]. The simplest way to create a parallel corpus in Sketch Engine is to upload data in a tabular format such as a spreadsheet (Excel). Spreadsheets must contain language names in the first row and then aligned segments (word, sentences, or paragraphs) side by side.

Every document (legal act) was processed to get one column for each language. In order to achieve this, bullets and numbering, as well as numerical part of tables in documents were removed. Each language in the source file is processed into a separate monolingual corpus and aligned with the corresponding corpus in the other language. It took approximately 30 man-days to perform this task.

Table 1. Size of parallel English-Croatian corpus

	English	Croatian
Tokens	1,151,297	1,059,406
Words	951,156	855,56
Sentences	72,045	75,638
Documents	178	178

[1] https://eur-lex.europa.eu/browse/directories/legislation.html

[2] https://www.sketchengine.eu/user-guide/user-manual/corpora/setting-up-parallel-corpora/

TERM EXTRACTION

Automatic term extraction is conducted for each language by using the option of extracting a list of 1000 single-word and multi-word keywords. The EUR-Lex English 2/2016 corpus was used as a reference corpus for extracting English single-word term candidates, while the English Web 2013 – term reference corpus was used as a reference corpus for extracting multi-word term candidates.

Similar options were possible for term extraction in Croatian. The EUR-Lex Croatian 2/2016 corpus served as a reference corpus for extracting more relevant single-word term candidates, while this option was not possible for extracting multi-term keywords. Thus the Croatian Web hrWaC 2.2. corpus was used instead.

In order to test the reliability of the method, the first 100 term candidates taken from multi-term keywords have been manually validated. Having checked their concordances, out of 100 term candidates, 70 was recognized as a valid term by a terminologist with ample experience in aviation terminology. The list of single-word keywords consisted largely out of aviation abbreviations and a few key aviation terms.

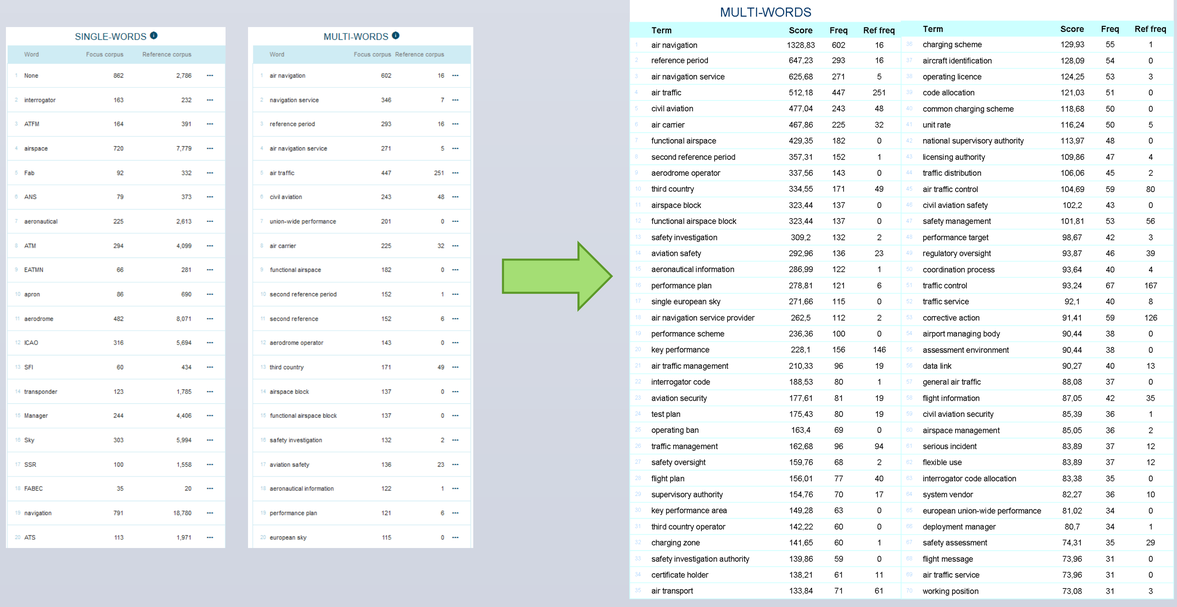

Figure 1. Automatically extracted single-word and multi-word keywords and a list of validated multi-word terms

This work has been fully supported by the Croatian Science Foundation under the project HRZZ-UIP-2017-05-7169.

loncarm@ihjj.hr aostrosk@ihjj.hr

TERM ANALYSIS

The concordances of most frequent English single-word and multi-word terms have been analyzed to check for their translation equivalents in Croatian. Although the majority of analyzed terms showed consistency in translation, larger constructions containing those terms – especially verbal constructions – provided good examples of potentially different use in different contexts.

Figure 2. Parallel concordances of the term *operate* and its translations in Croatian

Word sketches are automatic, corpus-derived summaries of words' grammatical and collocational patterns, and they serve as an excellent source for generating a larger number of term candidates in their syntactical context. Apart from typical adjective/noun (*wet-leased aircraft*) or noun/noun (*aircraft noise*, *aircraft position*) collocations, verbal collocations (*operate an aircraft*) and prepositional phrases (*aircraft entering into*) give invaluable terminological information most useful to translators.

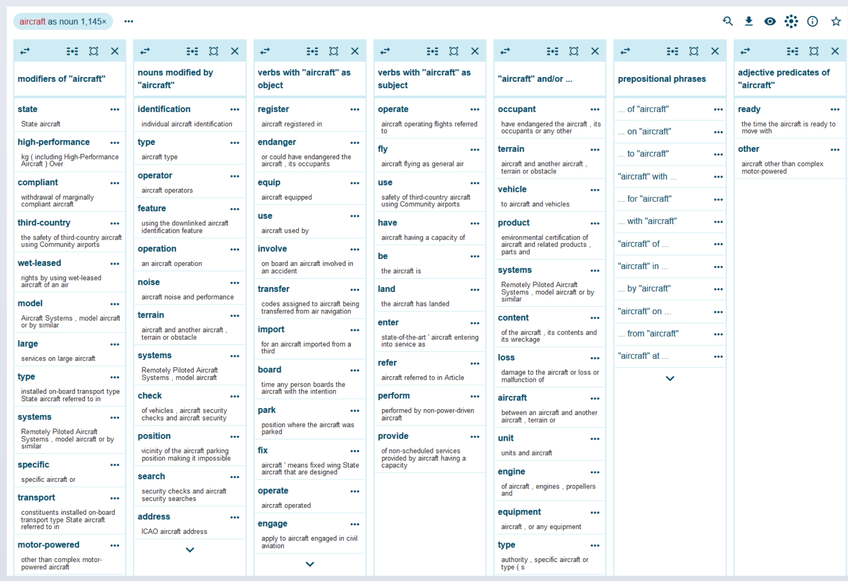

Figure 3. Word sketch of the term *aircraft*

CONCLUSION

The goal of this paper was to show how from publicly available texts in 24 official languages of the European Union with a freely available tool (for academic purposes) Sketch Engine one can assemble parallel corpora as a source of a valuable terminology information. The preliminary results of term analyses concerning the translation of key English terms in the field of air traffic provide evidence that a parallel corpus can be best used as the source of bilingual and multilingual terminology extraction and evaluation, intended specifically for building databases or LSP glossaries. Apart from the usual extraction of keywords or key terms in the field, parallel concordances can be used for semiautomatic definition extraction as well by using lexical markers such is shown in Figure 4. As is the case in monolingual specialized corpora, word sketches of single-word terms serve as a source of collocations and phraseological units, which are very often multi-word terms in themselves.

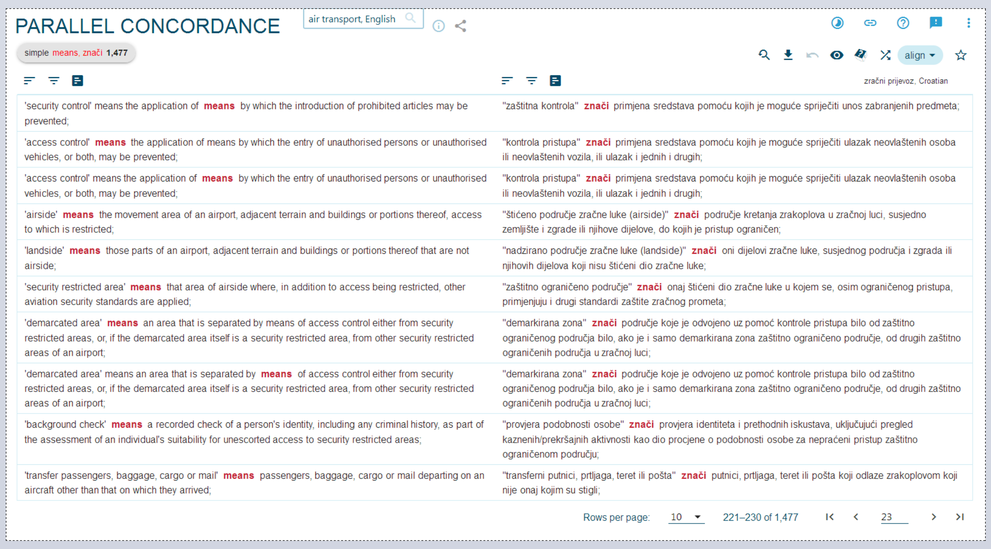

Figure 4. Definition extraction using lexical markers in parallel concordances