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 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

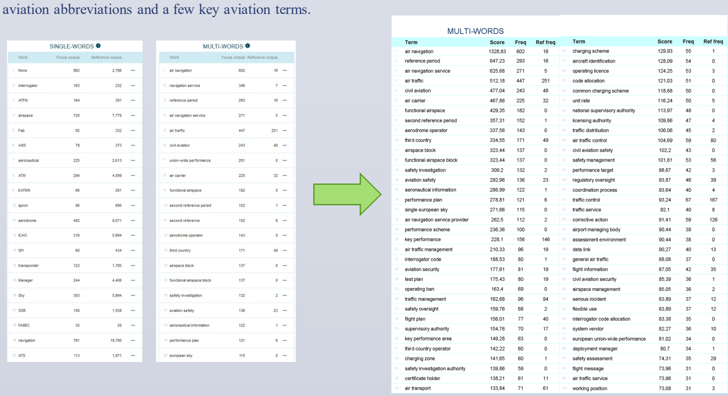
[2] https://www.sketchengine.eu/user-guide/user-manual/corpora/setting-up-parallel-corpora/

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.



SINGLE-WORDS

	Word	Focus corpus	Reference corpus	
1	None	862	2,786	···
2	interrogator	163	232	···
3	ATFM	164	391	···
4	airspace	720	7,779	···
5	Fab	92	332	···
6	ANS	79	373	···
7	aeronautical	225	2,613	···
8	ATM	294	4,099	···
9	EATMN	66	281	···
10	apron	86	690	···
11	aerodrome	482	8,071	···
12	ICAO	316	5,694	···
13	SFI	60	434	···
14	transponder	123	1,785	···
15	Manager	244	4,406	···
16	Sky	303	5,994	···
17	SSR	100	1,558	···
18	FABEC	35	20	···
19	navigation	791	18,780	···
20	ATS	113	1,971	···

MULTI-WORDS

	Word	Focus corpus	Reference corpus	
1	air navigation	602	16	···
2	navigation service	346	7	···
3	reference period	293	16	···
4	air navigation service	271	5	···
5	air traffic	447	251	···
6	civil aviation	243	48	···
7	union-wide performance	201	0	···
8	air carrier	225	32	···
9	functional airspace	182	0	···
10	second reference period	152	1	···
11	second reference	152	6	···
12	aerodrome operator	143	0	···
13	third country	171	49	···
14	airspace block	137	0	···
15	functional airspace block	137	0	···
16	safety investigation	132	2	···
17	aviation safety	136	23	···
18	aeronautical information	122	1	···
19	performance plan	121	6	···
20	european sky	115	0	···

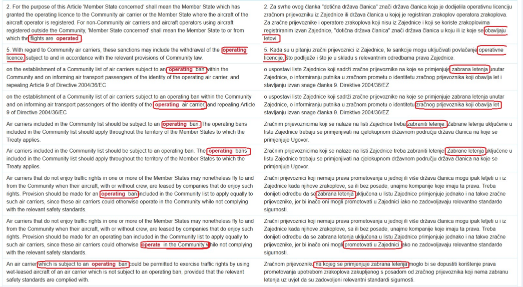
MULTI-WORDS

	Term	Score	Freq	Ref freq		Term	Score	Freq	Ref freq
1	air navigation	1328,83	602	16	36	charging scheme	129,93	55	1
2	reference period	647,23	293	16	37	aircraft identification	128,09	54	0
3	air navigation service	625,68	271	5	38	operating licence	124,25	53	3
4	air traffic	512,18	447	251	39	code allocation	121,03	51	0
5	civil aviation	477,04	243	48	40	common charging scheme	118,68	50	0
6	air carrier	467,86	225	32	41	unit rate	116,24	50	5
7	functional airspace	429,35	182	0	42	national supervisory authority	113,97	48	0
8	second reference period	357,31	152	1	43	licensing authority	109,86	47	4
9	aerodrome operator	337,56	143	0	44	traffic distribution	106,06	45	2
10	third country	334,55	171	49	45	air traffic control	104,69	59	80
11	airspace block	323,44	137	0	46	civil aviation safety	102,2	43	0
12	functional airspace block	323,44	137	0	47	safety management	101,81	53	56
13	safety investigation	309,2	132	2	48	performance target	98,67	42	3
14	aviation safety	292,96	136	23	49	regulatory oversight	93,87	46	39
15	aeronautical information	286,99	122	1	50	coordination process	93,64	40	4
16	performance plan	278,81	121	6	51	traffic control	93,24	67	167
17	single european sky	271,66	115	0	52	traffic service	92,1	40	8
18	air navigation service provider	262,5	112	2	53	corrective action	91,41	59	126
19	performance scheme	236,36	100	0	54	airport managing body	90,44	38	0
20	key performance	228,1	156	146	55	assessment environment	90,44	38	0
21	air traffic management	210,33	96	19	56	data link	90,27	40	13
22	interrogator code	188,53	80	1	57	general air traffic	88,08	37	0
23	aviation security	177,61	81	19	58	flight information	87,05	42	35
24	test plan	175,43	80	19	59	civil aviation security	85,39	36	1
25	operating ban	163,4	69	0	60	airspace management	85,05	36	2
26	traffic management	162,68	96	94	61	serious incident	83,89	37	12
27	safety oversight	159,76	68	2	62	flexible use	83,89	37	12
28	flight plan	156,01	77	40	63	interrogator code allocation	83,38	35	0
29	supervisory authority	154,76	70	17	64	system vendor	82,27	36	10
30	key performance area	149,28	63	0	65	european union-wide performance	81,02	34	0
31	third country operator	142,22	60	0	66	deployment manager	80,7	34	1
32	charging zone	141,65	60	1	67	safety assessment	74,31	35	29
33	safety investigation authority	139,86	59	0	68	flight message	73,96	31	0
34	certificate holder	138,21	61	11	69	air traffic service	73,96	31	0
35	air transport	133,84	71	61	70	working position	73,08	31	3

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

TERM ANALYSIS

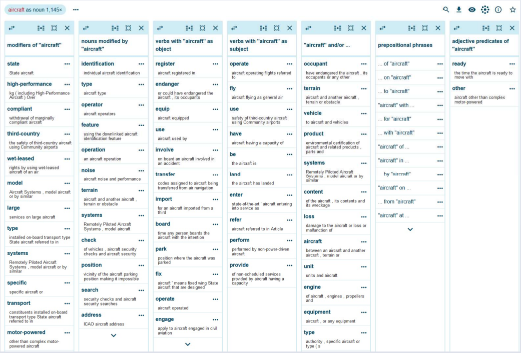
The concordances of most frequent English single-word and multi-word terms have been analyzed to check for their translation equivalents in Croatian. Although the majority of analyzed terms showed consistency in translation, larger constructions containing those terms – especially verbal constructions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



2. For the purpose of this Article 'Member State concerned' shall mean the Member State which has granted the operating licence to the Community air carrier or the Member State where the aircraft of the aircraft operator is registered. For non-Community air carriers and aircraft operators using aircraft registered outside the Community, 'Member State concerned' shall mean the Member State to or from which the flights are **operated** .	2. Za svrhe ovog članka "dotična država članica" znači država članica koja je dodijelila operativnu licenciju zračnom prijevozniku iz Zajednice ili država članica u kojoj je registriran zrakoplov operatora zrakoplova. Za zračne prijevoznike i operatore zrakoplova koji nisu iz Zajednice i koji se koriste zrakoplovima registriranim izvan Zajednice, "dotična država članica" znači država članica u koju ili iz koje se obavljaju letovi.
5. With regard to Community air carriers, these sanctions may include the withdrawal of the **operating** licence, subject to and in accordance with the relevant provisions of Community law.	5. Kada su u pitanju zračni prijevoznici iz Zajednice, te sankcije mogu uključivati povlačenje operativne licencije, što podliježe i što je u skladu s relevantnim odredbama prava Zajednice.
on the establishment of a Community list of air carriers subject to an **operating** ban within the Community and on informing air transport passengers of the identity of the operating air carrier, and repealing Article 9 of Directive 2004/36/EC	o uspostavi liste Zajednice koji sadrži zračne prijevoznike na koje se primjenjuje zabrana letenja unutar Zajednice, o informiranju putnika u zračnom prometu o identitetu zračnog prijevoznika koji obavlja let i stavljanju izvan snage članka 9. Direktive 2004/36/EZ
on the establishment of a Community list of air carriers subject to an operating ban within the Community and on informing air transport passengers of the identity of the **operating** air carrier, and repealing Article 9 of Directive 2004/36/EC	o uspostavi liste Zajednice koji sadrži zračne prijevoznike na koje se primjenjuje zabrana letenja unutar Zajednice, o informiranju putnika u zračnom prometu o identitetu zračnog prijevoznika koji obavlja let stavljanju izvan snage članka 9. Direktive 2004/36/EZ
Air carriers included in the Community list should be subject to an **operating** ban. The operating bans included in the Community list should apply throughout the territory of the Member States to which the Treaty applies.	Zračnim prijevoznicima koji se nalaze na listi Zajednice treba zabraniti letenje. Zabrane letenja uključene u listu Zajednice trebaju se primjenjivati na cjelokupnom državnom području država članica na koje se primjenjuje Ugovor.
Air carriers included in the Community list should be subject to an operating ban. The **operating** bans included in the Community list should apply throughout the territory of the Member States to which the Treaty applies.	Zračnim prijevoznicima koji se nalaze na listi Zajednice treba zabraniti letenje. Zabrane letenja uključene u listu Zajednice trebaju se primjenjivati na cjelokupnom državnom području država članica na koje se primjenjuje Ugovor.
Air carriers that do not enjoy traffic rights in one or more of the Member States may nonetheless fly to and from the Community when their aircraft, with or without crew, are leased by companies that do enjoy such rights. Provision should be made for an **operating** ban included in the Community list to apply equally to such air carriers, since these air carriers could otherwise operate in the Community while not complying with the relevant safety standards.	Zračni prijevoznici koji nemaju prava prometovanja u jednoj ili više država članica mogu ipak letjeti u i iz Zajednice kada njihove zrakoplove, sa ili bez posade, unajme kompanije koje imaju ta prava. Treba donijeti odredbu da se zabrana letenja uključena u listu Zajednice primjenjuje jednako i na takve zračne prijevoznike, jer bi inače oni mogli prometovati u Zajednici iako ne zadovoljavaju relevantne standarde sigurnosti.
Air carriers that do not enjoy traffic rights in one or more of the Member States may nonetheless fly to and from the Community when their aircraft, with or without crew, are leased by companies that do enjoy such rights. Provision should be made for an operating ban included in the Community list to apply equally to such air carriers, since these air carriers could otherwise **operate** in the Community while not complying with the relevant safety standards.	Zračni prijevoznici koji nemaju prava prometovanja u jednoj ili više država članica mogu ipak letjeti u i iz Zajednice kada njihove zrakoplove, sa ili bez posade, unajme kompanije koje imaju ta prava. Treba donijeti odredbu da se zabrana letenja uključena u listu Zajednice primjenjuje jednako i na takve zračne prijevoznike, jer bi inače oni mogli prometovati u Zajednici iako ne zadovoljavaju relevantne standarde sigurnosti.
An air carrier which is subject to an **operating** ban could be permitted to exercise traffic rights by using wet-leased aircraft of an air carrier which is not subject to an operating ban, provided that the relevant safety standards are complied with.	Zračnom prijevozniku na kojeg se primjenjuje zabrana letenja moglo bi se dopustiti korištenje prava prometovanja upotrebom zrakoplova zakupljenog s posadom od zračnog prijevoznika koji nema zabranu letenja uz uvjet da su zadovoljeni relevantni standardi sigurnosti.

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.



aircraft as noun 1,145×

modifiers of "aircraft"

- state: State aircraft
- high-performance: kg (including High-Performance Aircraft) Over
- compliant: withdrawal of marginally compliant aircraft
- third-country: the safety of third-country aircraft using Community airports
- wet-leased: rights by using wet-leased aircraft of an air
- model: Aircraft Systems , model aircraft or by similar
- large: services on large aircraft
- type: installed on-board transport type State aircraft referred to in
- systems: Remotely Piloted Aircraft Systems , model aircraft or by similar
- specific: specific aircraft or
- transport: constituents installed on-board transport type State aircraft referred to in
- motor-powered: other than complex motor-powered aircraft

nouns modified by "aircraft"

- identification: individual aircraft identification
- type: aircraft type
- operator: aircraft operators
- feature: using the downlinked aircraft identification feature
- operation: an aircraft operation
- noise: aircraft noise and performance
- terrain: aircraft and another aircraft , terrain or obstacle
- systems: Remotely Piloted Aircraft Systems , model aircraft
- check: of vehicles , aircraft security checks and aircraft security
- position: vicinity of the aircraft parking position making it impossible
- search: security checks and aircraft security searches
- address: ICAO aircraft address

verbs with "aircraft" as object

- register: aircraft registered in
- endanger: or could have endangered the aircraft , its occupants
- equip: aircraft equipped
- use: aircraft used by
- involve: on board an aircraft involved in an accident
- transfer: codes assigned to aircraft being transferred from air navigation
- import: for an aircraft imported from a third
- board: time any person boards the aircraft with the intention
- park: position where the aircraft was parked
- fix: aircraft ' means fixed wing State aircraft that are designed
- operate: aircraft operated
- engage: apply to aircraft engaged in civil aviation

verbs with "aircraft" as subject

- operate: aircraft operating flights referred to
- fly: aircraft flying as general air
- use: safety of third-country aircraft using Community airports
- have: aircraft having a capacity of
- be: the aircraft is
- land: the aircraft has landed
- enter: state-of-the-art ' aircraft entering into service as
- refer: aircraft referred to in Article
- perform: performed by non-power-driven aircraft
- provide: of non-scheduled services provided by aircraft having a capacity

"aircraft" and/or ...

- occupant: have endangered the aircraft , its occupants or any other
- terrain: aircraft and another aircraft , terrain or obstacle
- vehicle: to aircraft and vehicles
- product: environmental certification of aircraft and related products , parts and
- systems: Remotely Piloted Aircraft Systems , model aircraft or by similar
- content: of the aircraft , its contents and its wreckage
- loss: damage to the aircraft or loss or malfunction of
- aircraft: between an aircraft and another aircraft , terrain or
- unit: units and aircraft
- engine: of aircraft , engines , propellers and
- equipment: aircraft , or any equipment
- type: authority , specific aircraft or type (s

prepositional phrases

- ... of "aircraft"
- ... on "aircraft"
- ... to "aircraft"
- "aircraft" with ...
- ... for "aircraft"
- ... with "aircraft"
- "aircraft" of ...
- "aircraft" in ...
- by "aircraft"
- "aircraft" on ...
- ... from "aircraft"
- "aircraft" at ...

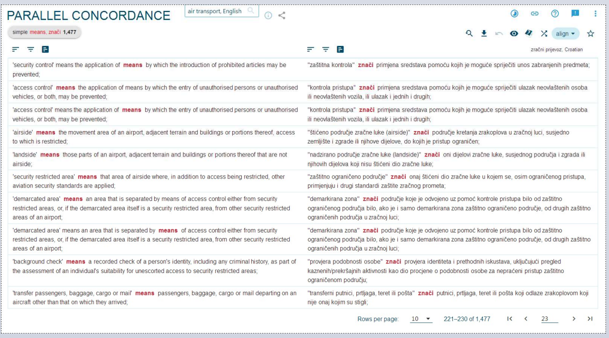
adjective predicates of "aircraft"

- ready: the time the aircraft is ready to move with
- other: aircraft other than complex motor-powered

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.



PARALLEL CONCORDANCE

air transport, English

simple means, znači 1,477

zračni prijevoz, Croatian

'security control' means the application of **means** by which the introduction of prohibited articles may be prevented;	"zaštitna kontrola" **znači** primjena sredstava pomoću kojih je moguće spriječiti unos zabranjenih predmeta;
'access control' **means** the application of means by which the entry of unauthorised persons or unauthorised vehicles, or both, may be prevented;	"kontrola pristupa" **znači** primjena sredstava pomoću kojih je moguće spriječiti ulazak neovlaštenih osoba ili neovlaštenih vozila, ili ulazak i jednih i drugih;
'access control' means the application of **means** by which the entry of unauthorised persons or unauthorised vehicles, or both, may be prevented;	"kontrola pristupa" **znači** primjena sredstava pomoću kojih je moguće spriječiti ulazak neovlaštenih osoba ili neovlaštenih vozila, ili ulazak i jednih i drugih;
'airside' **means** the movement area of an airport, adjacent terrain and buildings or portions thereof, access to which is restricted;	"štićeno područje zračne luke (airside)" **znači** područje kretanja zrakoplova u zračnoj luci, susjedno zemljište i zgrade ili njihove dijelove, do kojih je pristup ograničen;
'landside' **means** those parts of an airport, adjacent terrain and buildings or portions thereof that are not airside;	"nadzirano područje zračne luke (landside)" **znači** oni dijelovi zračne luke, susjednog područja i zgrada ili njihovih dijelova koji nisu štićeni dio zračne luke;
'security restricted area' **means** that area of airside where, in addition to access being restricted, other aviation security standards are applied;	"zaštitno ograničeno područje" **znači** onaj štićeni dio zračne luke u kojem se, osim ograničenog pristupa, primjenjuju i drugi standardi zaštite zračnog prometa;
'demarcated area' **means** an area that is separated by means of access control either from security restricted areas, or, if the demarcated area itself is a security restricted area, from other security restricted areas of an airport;	"demarkirana zona" **znači** područje koje je odvojeno uz pomoć kontrole pristupa bilo od zaštitno ograničenog područja bilo, ako je i samo demarkirana zona zaštitno ograničeno područje, od drugih zaštitno ograničenih područja u zračnoj luci;
'demarcated area' means an area that is separated by **means** of access control either from security restricted areas, or, if the demarcated area itself is a security restricted area, from other security restricted areas of an airport;	"demarkirana zona" **znači** područje koje je odvojeno uz pomoć kontrole pristupa bilo od zaštitno ograničenog područja bilo, ako je i samo demarkirana zona zaštitno ograničeno područje, od drugih zaštitno ograničenih područja u zračnoj luci;
'background check' **means** a recorded check of a person's identity, including any criminal history, as part of the assessment of an individual's suitability for unescorted access to security restricted areas;	"provjera podobnosti osobe" **znači** provjera identiteta i prethodnih iskustava, uključujući pregled kaznenih/prekršajnih aktivnosti kao dio procjene o podobnosti osobe za nepraćeni pristup zaštitno ograničenom području;
'transfer passengers, baggage, cargo or mail' **means** passengers, baggage, cargo or mail departing on an aircraft other than that on which they arrived;	"transferni putnici, prtljaga, teret ili pošta" **znači** putnici, prtljaga, teret ili pošta koji odlaze zrakoplovom koji nije onaj kojim su stigli;

Figure 4. Definition extraction using lexical markers in parallel concordances

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