UDK 811.163.42:811.521'374
379.8:910.4
Izvorni znanstveni rad
Rukopis primljen 20. XI. 2019.
Prihvaćen za tisak 23. I. 2020
doi.org/10.31724/rihjj.46.2.31

Irena Srdanović
Juraj Dobrila University of Pula
Zagrebačka ulica 30, HR-52100 Pula
irena.srdanovic@unipu.hr

FROM SPECIALIZED WEB CORPORA OF TOURISM TO A LEARNER'S DICTIONARY

This paper presents the two approaches used in creating specialized web corpora of Croatian tourism in Japanese for their usage in building a specialized learners' dictionary. Both approaches use the WebBootCat technology (Baroni et al. 2006, Kilgarriff et al. 2014) to automatically create specialized web corpora. The first approach creates the corpora from the selected seed words most relevant to the topic. The second approach specifies a number of web pages that cover tourism-oriented information on specified regions, cities, and sites in Croatia available in Japanese, which are then used for web corpora creation inside the Sketch Engine platform. Both approaches provide specialized web corpora small in size, but quite useful for lexical profiling in the specific field of tourism. In the process of dictionary creation, the second approach has proven to be especially useful for the selection of lexical items, while both approaches have proven to be highly useful for the exploration and selection of authentic examples from the corpora. The research exposes some shortcomings in Japanese language processing, such as errors in the lemmatization of some culturally specific terms and indicates the need to refine existing language processing tools in Japanese. The Japanese-Croatian bilingual learner's dictionary (Srdanović 2018) is currently in the pilot phase and is being used and built by learners and teachers through the open-source dictionary platform Lexonomy (Mechura 2017). In addition to the fact that work on the bilingual dictionary is useful as a means for training students in language analysis and description using modern technologies (e.g. corpora, corpus query systems, dictionary editing platform), the dictionary is also important in educating new personnel capable of working in tourism using the Japanese language, which is strongly needed. In future, the same approach could be used for creating specialized corpora and dictionaries for Japanese and other language pairs.

1. Introduction

Due to the growing number of tourists from Japan and Asia, there is an increasing need for Japanese-speaking guides and other personnel in the Croatian tourism industry. Taking that into consideration, the Japanese language and culture study program at Juraj Dobrila University of Pula offers courses on the Japanese language in tourism in addition to general courses in Japanese language and culture. The courses provide tourism-oriented specialized language learning to assist students in acquiring the skills and specialized vocabulary needed to prepare students for future jobs with specialized Japanese language skills in the tourism industry. To support and respond to these educational needs among the growing number of students of the Japanese language and culture, keeping them in line with demands from the market and society, it is important to work continuously on developing various language learning resources in the students' native language (i.e. bilingual dictionaries, textbooks, grammars, etc.). While there are plenty of English-Japanese and Japanese-English resources, resources in Japanese-Croatian and Croatian-Japanese language pairs or even in pairs with other closely related Slavic languages, such as Serbian or Slovene, are very limited (e.g. Yamasaki-Vukelić 2006, Erjavec et al. 2006, Hmeljak Sangawa and Erjavec 2010). Therefore, the Department of Asian studies at the Juraj Dobrila University of Pula has made initial efforts in developing resources, such as specialized corpora, dictionaries, textbooks, while using the latest technologies such as corpus query systems, corpus builders, online dictionary platforms, etc. Due to a strong need for a learner's dictionary with more detailed information, such as frequent patterns, examples of word usage, user notes and difficulty levels, a few pilot dictionary projects have been initiated, targeting students not only as users, but also as co-creators of the dictionaries (Srdanović 2018a, 2018b). The project also recognizes the need for more resources on Croatian tourism in Japanese. The need for such resources exists primarily among Japanese language learners who aim to become future personnel in the tourism industry, but also for local services, as well as for Japanese speaking tourists.

This paper explores two methods used in creating a specialized web corpus of Croatian tourism in the Japanese language using the WebBootCat technology within the Sketch Engine platform for the automatic building of specialized web corpora (Baroni et al. 2006, Kilgarriff et al. 2014). The first approach creates the

corpora from selected seed words most relevant to the topic, while the second approach uses a number of web pages that cover tourism-oriented information for specific regions, cities, and sites in Croatia available in Japanese to create specialized corpora. The students are involved in the process of creating the corpora, to use them as a source for extracting keywords and examples and building a bilingual Japanese-Croatian online learner's dictionary of terminology related to Croatian tourism.

In the following chapter, I will examine recent technology for building corpora and dictionaries, mainly WebBootCat technology, Sketch Engine, and Lexonomy, recent research involving technologies, and describe how these technologies were introduced to students of Japanese. In the third chapter, methods used in this study for creating a specialized web corpus of vocabulary on Croatian tourism in Japanese are described, with a comparison of the results from the perspective of dictionary building.

2. Technologies for building corpora and dictionaries

2.1. WebBootCaT technology

The BootCat technology was initiated by Baroni and Bernardini (2004) as a method for bootstrapping corpora and terms from the web and instant corpus building for a specific domain. It compares frequencies in specialized and reference corpora to look for terms typical of the specialized data. The technology uses the following steps:

1) The user inputs a few domain-specific seed words – these seed words must be specific for the domain, and may not appear in any other domain to avoid the collection of data not relevant to the targeted domain;

2) The seed words are sent in various combinations to a search engine; four random seed words must appear on one web page to be identified;

3) The hits identified by the search engine are gathered, cleaned, de-duplicated, and processed to give a domain-specific corpus.

Baroni and Bernardini (2004) used the BootCat technology to build English and Italian corpora and term lists from the domain of psychiatry. Later on, this tech-

nology was applied to create various small specialist corpora for finding terminology and translations (e.g. Baroni and Ueyama 2004) and also large, general ones (e.g. Sharoff 2006, Srdanović et al. 2008). Kilgarriff (2014) included the technology into the Sketch Engine platform, described its functionality and possibility to find terms in corpora for many languages. The method manages to compare the domain-specific corpus to a reference corpus and to provide keywords and terms for a specific domain. The automatic term recognition (ATR) process proposes term candidates, which are then accepted or rejected by the user.

Furthermore, Kilgarriff (2009) introduced the possibility of using the WebBootCat to build corpora on the topic of the student's own choice for language learning purposes. The direct use of corpora in the classroom has proven to be challenging due to the complexity of concordances and the difficulty in reading them, but an alternative strategy is to use them in the form of an automatic collocations 'dictionary' built into the Sketch Engine. This can be achieved by using word sketch techniques (described below) to present to the students the needed information not available in dictionaries in a more digested way.

This study is the first attempt to create a domain-specific corpus in the Japanese language targeted at tourism-related terminology for the Croatian region. Besides, this is the first attempt to compare two methods and use the data for bilingual lexicography for these two language pairs.

2.2. Sketch Engine

The Sketch Engine (SkE) (Kilgarriff et al. 2004) is a state-of-the-art corpus query system with various functionalities, such as building and managing corpora, searching lexical items using concordances, word sketches – summary of lexical profile of words, comparing word sketches, thesaurus, etc. It has been used for around 150 languages, among which Japanese. Srdanović et al. (2008, 2012) have prepared the system for use in Japanese, and at this stage, it covers part of speech tagging, word sketches grammar, lemmatization, and term extraction. Below, we briefly introduce some of the functionalities within SkE that are most relevant for this paper.[1]

[1] Sketch Engine User Guide (www.sketchengine.eu/guide) provides detailed information on the functionalities.

The Sketch Engine enables corpus building using either available files or the web. Building a corpus from relevant web pages can be done using the above described WebBootCaT technology, by providing some typical words defining the topic (seed words), providing a list of URLs or by downloading a complete website. This study explores the first two methods.

The Concordance functionality enables searching the corpus in a variety of ways for words, phrases, tags, documents, text types, or corpus structures. The search results are displayed in context and can be further sorted and processed. The CQL (Corpus Query Language) search enables complex searches with regular expressions, tags, lemmas, words, etc.

Word Sketches is a one-page summary of the grammatical and collocational behavior of searched words. They are organized into grammatical relations, which have been previously defined by rules written in the word sketches grammar for a particular language.

Keywords and term extraction are used to identify typical words for a specific domain mainly for use in translation and interpreting, to identify what is unique in the first corpus compared to the second one, or to define the content or topic typical for a specific text. The results are divided into keywords (single word items) and terms (multiword items) and are based on the comparison of the focus corpus to a reference one in the language in question.

GDEX (Good Dictionary EXamples) automatically selects examples from the corpora with respect to their suitability to serve as dictionary examples. (Kilgarriff et al. 2008; for Japanese GDEX refer to Srdanović and Kosem 2016, Srdanović 2019).

2.3. Lexonomy

Lexonomy[2] (Mechura 2017) is a cloud-based online open-source platform for creating, editing, and publishing dictionaries online. It is freely available for users and provides a short and user-frendly introduction to the tool for complete beginners. It can be used by multiple users, which also makes it convenient for

[2] www.lexonomy.eu

working in a team of teachers and students. The dictionary tool provides two simple templates for monolingual and bilingual dictionaries, but dictionary creation without a template is also an option. Entries in the dictionary have previously defined the structure of elements and are stored as XML (Extensible Markup Language) documents. The configuration interface allows for the definition of the internal structure of the dictionary, the way the elements are formatted, as well as user permissions and the external appearance of the dictionary. The platform is used for the ongoing project Japanese-Croatian learners' dictionary, including the dictionary specialized for the domain of tourism.

2.4. Introducing novel technologies to students

The Introduction to Lexicology and Lexicography course aims to define and explain the basic principles of Japanese and general lexicology and lexicography and apply that knowledge to the creation of the dictionary and analysis using various novel technologies. The third-year students of the Japanese language and culture program enrolled in the course during academic years 2017/18 and 2018/19[3], which lasted 15 weeks (one semester), 90 minutes per week, and provided 4 ECTS.

After a gentle introduction to corpus lexicography and various Japanese language corpora, the Sketch Engine and other corpus query systems are demonstrated to the students. The students receive some assignments to become familiar with the corpora and tools. The WebBootCat technology, Word sketches, Comparing word sketches, Thesaurus, and other functionalities within the Sketch Engine are demonstrated to the students, who are encouraged to use them for various activities and instructed to perform various tasks as independent work. One of the tasks was to build a web corpus of vocabulary concerning Croatian tourism in Japanese, which is described in the following section. Finally, the dictionary-editing platform Lexonomy is demonstrated to the students followed by a team project on planning and configuring the structure of the target Japanese-Croatian learners' dictionary, which is then defined to be used consistently for the duration of the dictionary project (Srdanović 2018a, 2018b).

[3] From the academic year 2019/2020 onward, the course was relocated to the master's program in Japanese studies.

3. Building specialized corpora of Croatian tourism in Japanese

3.1. The first method: seed words

The first method used for building the specialized corpora of Croatian tourism vocabulary in Japanese was based on selected seed words. The students were instructed as follows:

1. Prepare a list of 20 words/phrases (seed words) in Japanese that are related to Croatian tourism – Croatian town names or some sights, typical Croatian dishes, etc. in Japanese, e. g. the Japanese words for Zagreb (ザグレブ) or štrukle (シュトゥルクリ). These words as seed words are expected to automatically collect web pages on Croatian tourism.

2. Make sure that these words / phrases can be found on the Internet in Japanese and are not used in other domains/languages or are not related to other languages/cultures as well. For example, the Japanese word for Istrian pasta *fuži* is フジ but this string of letters can be easily found as other words in Japanese language data, such as the mountain Fuji written in katakana, the company name *Fuji* Company, Limited (株式会社フジ), the television station name *Fuji* TV (フジテレビ), which would cause collecting unwanted pages.[4]

3. In the Sketch Engine web tool, use the WebBootCat functionality to create a web corpus (Figure 1). The tool uses your seed words to automatically collect web sites that include combinations of four random seed words from the list of 20.

4. When the corpus is prepared, try different functions in Sketch Engine tool to explore the results (Word List, Collocations, Frequency, Concordance Query, Word Sketch, etc.).

[4] Another example to avoid as a seed term is the word for the Croatian town Pula, in Japanese プーラ (Puura), which can also refer to a tool or a hair care product in Japanese. Furthermore, the word *sarma* (サルマ) could be a good seed term for a corpus on Croatian tourism as one of typical dishes in Croatia, but since it is also typically eaten in other regions on the Balkan Peninsula, it would as a seed term collect web pages related to tourism in wider regions. Therefore, it needs to be avoided as a seed term for the type of corpus targeted in the current research, but could be used in the future as a seed term for building a wider corpus targeted at Balkan region; the term itself is, however, expected to appear in the final corpus.

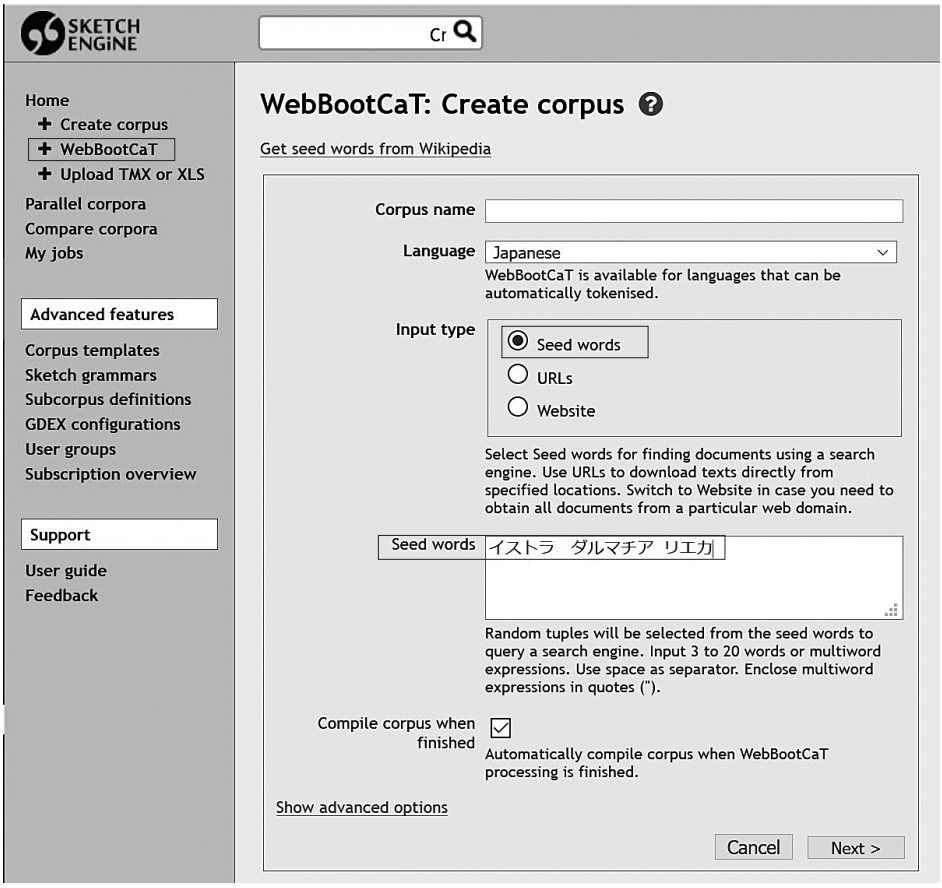


Figure 1: The WebBootCat page in Sketch Engine with sample seed words

Figure 2 shows the scheme of the procedure: Using seed words, we create a web corpus, which is then used to explore keywords, terms, collocations, concordances, examples, etc.

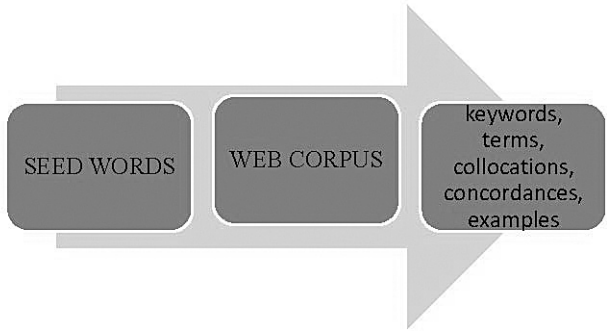


Figure 2: The first method (seed words) scheme

3.2. Results: the first method (seed words)[5]

Several small web corpora were created in the domain of tourism: Croatian tourism in Japanese, with sizes from between 1 million and 12 million tokens. Table 1 presents statistical information regarding the created corpora, with a few examples of seed words in Japanese and their translations into English. Croatian Tourism Corpus 1 is oriented towards seed words from the Istrian peninsula and peculiarities of the region covering various towns, tourist sites, but also including traditional Istrian food and drinks. Croatian tourism corpora 2 and 3 cover various towns and tourism sites from all over Croatia, the only difference being that corpus 2 encompasses more general and overall notions, such as *Adriatic Sea*, *Dalmatia*, while corpus 3 encompasses more complex terms, such as *Zagreb's licitar heart*, which most likely rarely retrieve results. This must be the main reason why corpus 2 is larger than corpus 3. Corpus 4 concentrates on the town Rovinj, with various specifics of the town, but also includes general terms, such as *Croatia*. It is to be expected that the size of the corpus is related to the type of seed words chosen. In cases of more general terminology targeted at the Croatian region overall, more data is retrieved than in the case of specific terms

[5] In the SkE terminology, "keywords (single words)" refer to "individual words (tokens) which appear more frequently in the focus corpus than in the reference corpus", while "terms (multi-words)" refer to "multi-word expressions which appear more frequently in the focus corpus than in the reference corpus and, additionally, match the typical format of terminology in the language" (www.sketchengine.eu). The reference corpus for Japanese is JaTenTen11, LUW (Japanese web sample corpus created in 2011 and it is annotated with UniDic's long-unit word units).

targeted at smaller regions. Also, very complex terms consisting of multiword expressions lower the possibility of including more pages resulting in smaller sized corpora with fewer tokens and words.

Although the students were instructed not to use the terms used in other domains/languages or terms that are related to other languages/cultures (see Section 3.1: step 2) as seed words, such terms appear a few times (e. g. サルマ 'sarma' [traditional food in the Balkans] – used in the wider region, レース reesu 'lace, race' – used in different domains).

Qualitative screening of raw corpus data, single-words (keywords), and multiword units (terms) reveals that the results of method 1 are not satisfactory and need further elaboration. While we can sporadically find keywords and terms related to Croatian tourism in Japanese, such as the names of tourist destinations and sites (アドリア海 *Adoria-kai* 'Adriatic Sea', ザグレブ *Zagurebu* 'Zagreb'), descriptions (e. g. 有利な陣地 *yuurina jinchi* 'advantageous position'), terms for periods (紀元前八世紀 *kigenzen hachi seiki* 'BC 8th century') or general tourist terminology, such as 文化遺産 *bunka isan* 'cultural heritage', 地図 *chizu* 'map', there is still a large amount of noise in the data. The noise can be classified as follows: alphabetical letters and words, numerals with or without combinations with words, dates, words from other domains, e.g. music: 弦楽 *gengaku* 'string music', 歌劇 *kageki* 'opera', 交響曲 第1番 *koukyoukyoku daiichiban* 'Symphony No.1' or race: オートバイ レーサー *ootobai reesaa* 'motorcycle racer', nonsense words, etc.

Table 1: Croatian tourism corpora seed words and statistics

	Seed word examples	Documents	Words	Tokens	Sentences
Croatian tourism corpus 1 (Istria)	イストラ半島 Isutora hantou 'the Istrian peninsula', アウグストゥス神殿 Augusutousu shinden 'Temple of Augustus', ウチュカ自然公園 Uchuka shizen kouen 'National Park Učka', モトヴン Motovun 'Motovun', イストラ牛 Isutora gyuu 'Istrian cattle', メディツァ meditsa 'medica [drink]'	76	1,150,326	1,788,501	106,438
Croatian tourism corpus 2 (all over Croatia)	アドリア海 Adoria-kai 'the Adriatic Sea', ダルマチア Darumatia 'Dalmatia' スプリット Supritto 'Split' レポグラヴァのレース Repogurava no reesu 'Lepoglava lace', ザグレブ Zagurebu 'Zagreb', イェラチッチ広場 Ierachicchi hiroba 'Ban Josip Jelačić Square', クレームシュニッタ Kureemushunitta 'kremšnita'	547	5,928,991	9,370,268	591,803

Croatian tourism corpus 3 (all over Croatia)	ザグレブ Zagurebu ‘Zagreb’, レース reesu ‘lace’, プリトヴィッツェ Puritowittse ‘Plitvice’, オシエク Oshieku ‘Osijek’, ディオクレティアヌス宮殿 Diokurechianusu kyuuden ‘Diocletian’s Palace’, “ザグレブの赤いハート「リツィタル」” Zagurebu no akai haato ritsitaru ‘Zagreb’s licitar heart’	373	3,632,694	5,906,681	302,183
Croatian tourism corpus 4 (Rovinj, Pula, Croatia)	バタナ漁船 Batana gyosen ‘Batana Fishing Boat’, リム湾 Rimu-wan ‘Lim Bay’, モンコドニャ Monkodonya ‘Moncodonya’, クロアチア Kuraochia ‘Croatia’, バルビ門 Barubi-mon ‘Barbi Gate’, ロヴィニ Rovini ‘Rovinj’, ブロデット burodetto ‘brudet’, ブザラ buzara ‘buzara’	426	8,179,722	12,091,918	636,657

Furthermore, a problem in the Japanese language annotation of foreign-culture related terms has been discovered. Figure 3 shows Japanese language concordances for ダルマチア Darumachia ‘the Dalmatia region’, but in keyword extraction, where the lemma is withdrawn, it can be noted that this word is incorrectly lemmatized as 達磨チアー Darumachiaa consisting of the words Daruma ‘a Japanese traditional doll that symbolizes perseverance and good luck’ and chiaa ‘cheers’. This occurs because Japanese morphological analyzer MeCab and the UniDic dictionary used for annotation do not recognize this specific terminology.

The following steps are needed to further clean-up the data: checking seed words, checking and removing corpus files, and resolving possible technical issues in the process specific to the Japanese language.

Figure 3: Concordances for Japanese words concerning the region of Dalmatia in Croatian tourism corpus 2

3.3. The second method: URLs collection

The second method for building specialized corpora of Croatian tourism in Japanese consisted of collecting a list of Japanese web pages related to Croatian tourism. The following instructions were given to the students:

1. Choose a region in Croatia, e.g. Istria.
2. Explore existing Japanese words for that region, cities, and tourist attractions, e. g. イストラ *Isutora* 'Istria', プ―ラ *Puura* 'Pula', プーラ円形劇場 *Puura enkei gekijou* 'The Pula Amphitheater, Arena'.
3. Search for these Japanese words on the Internet to find information on the region, cities, and tourist attractions using these Japanese words as relating to tourism.
4. Compile a list of all visited URL sites with Japanese texts on Croatian tourism.

After completing the individual tasks of collecting URL sites, all of the collected URL sites were assembled to automatically build the corpus within SkE, as shown in Figure 4.

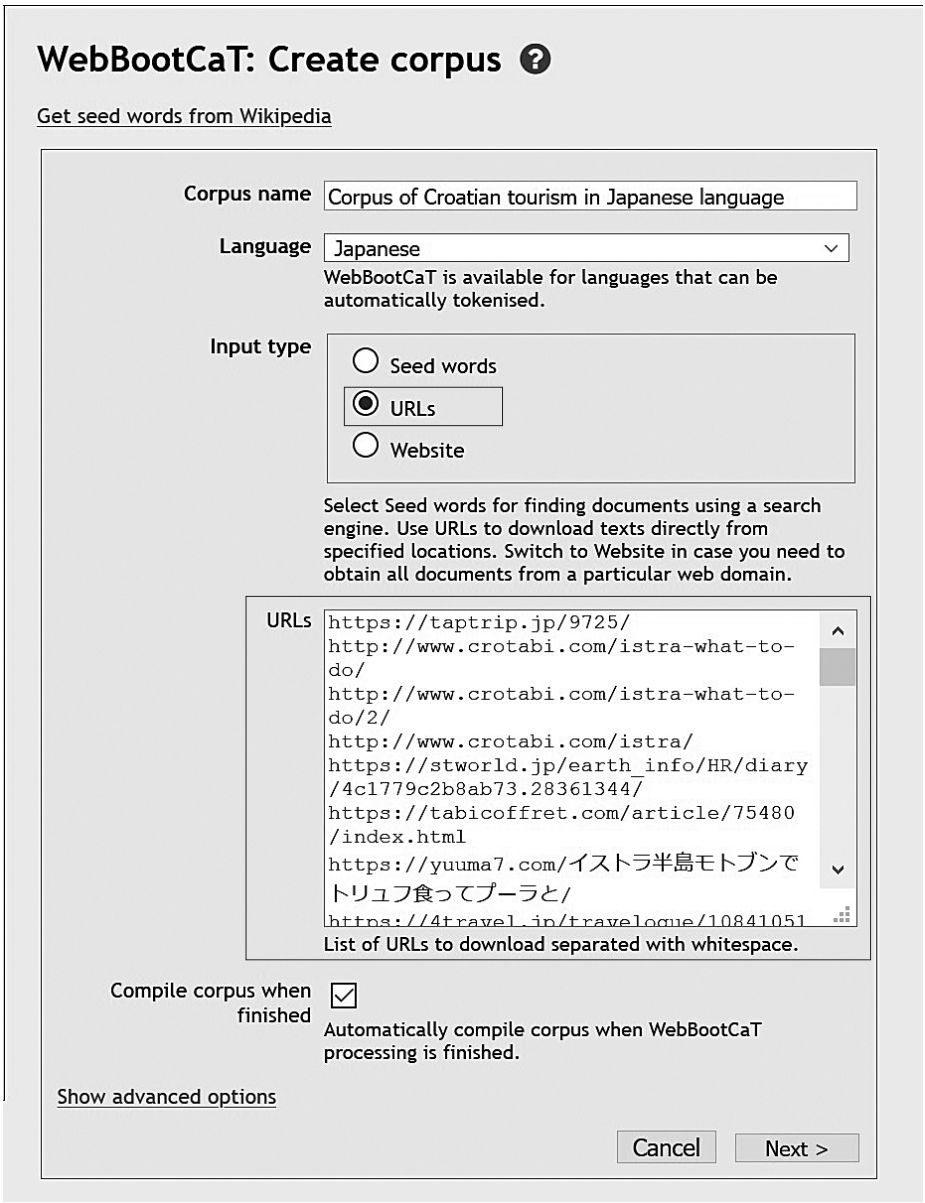


Figure 4: The WebBootCat page in Sketch Engine with URLs

Figure 5 shows the scheme of the procedure: creating a web corpus using URL sites, which is then used to explore keywords, terms, collocations, concordances, examples, etc.

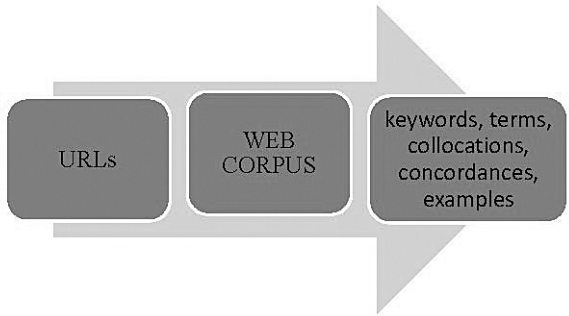


Figure 5: The second method (URLs) scheme

3.4. Results: the second method (URLs)

The second method resulted in 150 web sites in Japanese covering the domain of Croatian tourism collected, upon which a small-sized web corpus (around 130,000 tokens), the Corpus of Croatian Tourism in Japanese, was created. Table 2 presents statistics on the corpus.

Table 2: Statistics on the Corpus of Croatian Tourism in Japanese (URLs)

Counts		**Common tags**		**Lexicon sizes**	
Tokens	129,256	noun	N.*	Word	18,859
Words	105,312	verb	V.*	Lemma	15,817
Sentences	8,806	adjective	Ai.*	lemma_kana	16,845
Paragraphs	2,730	adverb	Adv.*	Tag	48
Documents	98	pronoun	Pron	infl_type	61
		conjunction	Conj	infl_form	27
		particle	P.*		
		interjection	Interj		

Figure 6 presents the first ten keywords (single-words) from the corpus along with their frequency and relative frequency in the focus corpus (Corpus of Croatian tourism in Japanese) in comparison with their appereance in the reference corpus (JaTenTen11, LUW sample). The keywords are ordered based on the keyness score calculated using simple math statistics, a method of identifying the keywords of one corpus vs. another, implemented within the Sketch Engine tool. The method compares the normalized (per million) frequency of the

word in the focus corpus vs. the normalized (per million) frequency of the word in the reference corpus (Kilgarriff 2009b).[6] All of the keywords with the high keyness score are actually a good representative of Croatian tourism: クロアチア *Kuroachia* 'Croatia', ドゥブロヴニク *Doburovuniku* 'Dubrovnik', ザグレブ *Zagurebu* 'Zagreb', アドリア海 *Adoriakai* 'the Adriatic Sea', イストラ半島*Isutora hantou* 'the Istrian peninsula', リエカ *Rieka* 'Rijeka', 旧市街 *kyuushigai* 'old town', オパティヤ *Opatia* 'Opatia', ドブロブニク *Doburobuniku* 'Dubrovnik', コルチュラ島 *Koruchura-tou* 'the island of Korčula'. In addition to the name of the country, the capital, and other towns and regions, among the first ten words, we find 旧市街 *kyuushigai* 'the old town', a general term that is often used in tourist descriptions.

It should be mentioned that the keywords are selected as *words* to achieve the above results since it results in better output than the keywords as *lemmas* (the default setting of the keyword extraction functionality within the SkE). In the case of lemmas, the results fail to show accurately some words, as Japanese processing tools wrongly annotate them. For example, the aforementioned 達磨チア一 *Daruma chiaa* '[sig.] Daruma + cheer' instead of 'Dalmatia' and 椅子虎半島 *Isu tora hantou* '[sig.] chair+tiger+peninsula' instead of 'the Istrian peninsula' appear among the first ten lemmas. Opening the corpus examples of these lemma keywords reveals the real usage of the terms as words ダルマチア *Darumachia* 'Dalmatia' and イストラ半島 *Isutora hantou* 'the Istrian peninsula'.

[6] For more information on statistics within SkE, refer to www.sketchengine.eu/documentation/statistics-used-in-sketch-engine.

SINGLE-WORDS MULTI-WORDS ABOUT

reference corpus: Japanese Web 2011 sample (jaTenTen11, LUW)

	Word	Frequency[?] Focus	Frequency[?] Reference	Relative freq.[?] Focus	Relative freq.[?] Reference	Score [?]
1	クロアチア	445	310	3,442.78	1.522	1,365.48
2	ドゥブロヴニク	133	2	1,028.966	0.01	1,019.95
3	ザグレブ	145	33	1,121.805	0.162	966.25
4	アドリア海	137	63	1,059.912	0.309	810.28
5	イストラ半島	92	2	711.766	0.01	705.83
6	リエカ	84	3	649.873	0.015	641.43
7	旧市街	173	295	1,338.429	1.448	547.07
8	オパティヤ	68	1	526.088	0.005	524.51
9	ドブロブニク	69	20	533.824	0.098	487
10	コルチュラ島	61	0	471.932	0	472.93

Figure 6: The first ten keywords in the Corpus of Croatian tourism in Japanese (compared to JaTenTen, LUW)

MULTI-WORDS

	Word	Focus corpus [?]	Reference corpus [?]
1	2012年 2月号	23	0
2	小高い 丘	12	145
3	美しい ビーチ	7	23
4	美しい 景色	7	99
5	美しい 海	7	113
6	美しい 町	6	41
7	ロマンチック な 街	6	1
8	ロマンチック な 雰囲気	6	13
9	美しい 風景	6	141

Figure 7: The first multi-word units in the Corpus of Croatian tourism in Japanese (compared to JaTenTen, LUW)

Furthermore, Figure 7 shows the multi-word units with the highest keyness score in the corpus. The majority of the results are terms representative of Croatian tourism in Japanese, such as 小高い丘 *kodakai oka* 'a small hill', 美しいビーチ *utsukushii biichi* 'beautiful beaches', 美しい景色 *utsukushii keishiki* 'a beautiful view', 美しい海 *utsukushii umi* 'a beautiful sea', *utsukushii machi* 'a beautiful town', ロマンチックな町 *romanchikkuna machi* 'a romantic town', ロマンチックな雰囲気 *romanchikkuna fun'iki* 'romantic atmosphere', 美しい風景 *utsukushii fuukei* 'beautiful scenery'. However, there is a need to exclude some of the data from the results, such as 2012年2月号 2012 *nen 2gatsu-gou* 'issue Feb 2012 (newsletter)', the number of stays, hours, various types of quantities (e. g. height), incomplete multi-word units, such as 多い冬 *ooi fuyu* 'winter with a lot of …' which appears as, for example, 雨の多い冬 *ame no ooi fuyu* 'winter with a lot of rain', functional words with no specific meaning, etc.

Finally, the lists of 200 keywords and 200 multi-word units were evaluated for their relevance as terminology representative of Croatian tourism in Japanese and their relevance for inclusion in the Japanese-Croatian learner's dictionary of tourism. The results reveal that 96% of extracted keywords are related to tourism in Croatia, among which 44% belong to general terms of tourism, 40% to geographical names and 12% to touristic sites, while only 4% were inadequate (for example, symbols, letters, words written in the alphabet, etc.). As for multiword expressions, 68% were expressions related to tourism that could be included in a dictionary, and 32% were inadequate, as they represent various dates, periods, durations, or are too general or too specific. The overall results show that the corpus data is of high quality and very usable for the target dictionary. Moreover, this corpus is also a good source for dictionary examples. However, there is a need to enlarge the collection and the corpus to obtain more data, as well as to overcome the issues with language processing errors of a foreign language and foreign culture-specific terms.

3.5. Comparison of corpora results and their possible use for the dictionary

Using the two methods described above, we have created four new Croatian tourism corpora (using method 1: seed words) and the new Corpus of Croatian tourism in Japanese (using method 2: URL sites). An initial scan of newly cre-

ated corpora has revealed that the second method has given more satisfactory results in terms of the quality of keywords and multi-word units, which can be used as entries for the bilingual dictionary. The evaluation of keywords and multi-word units in the corpus confirmed the quality of the data. However, the corpus is rather small and needs to be enlarged.

To further evaluate and compare the corpora results, we searched random Japanese terms related to Croatian tourism, one from each group of the seed words throughout all new Croatian tourism corpora, plus an additional word for the capital. The searched terms are: アウグストゥス神殿 *Augustus shinden* 'Augustus' temple', オシエク *Oshieku* 'Osijek, town', プリトヴィッツェ *Puritovittse* 'Plitvice, National Park', バルビ門 *Barubi-mon* 'Barbi Gate'. The values from the new specialized corpora are then compared to the values from the Japanese reference web corpus JaTenTen2011, LUW, in frequency and in normalized frequency for the size of corpora (per million). The results are presented in Table 3 (all the numerals for the terms that were used as seed words in a specific corpus are marked grey).

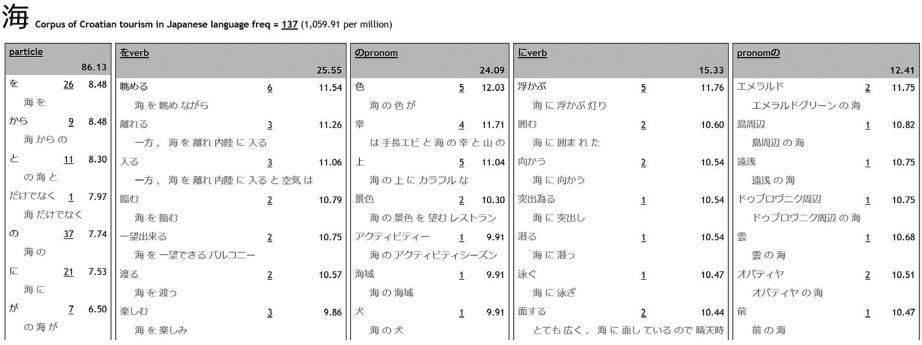
Table 3: Compared corpora for searched terms on Croatian tourism

		アウグストゥス神殿 Augustus shinden 'Augustus' temple'	オシエク Oshieku 'Osijek, town'	プリトヴィッツェ Puritovittse 'Plitvice National Park'	バルビ門 Barubi-mon 'Barbi Gate'	ザグレブ Zagurebu 'Zagreb'
Croatian tourism corpus 1 (Istria)	Freq	14	0	1	0	41
	per million	7.83	0	0.56	0	22.9
Croatian tourism corpus 2 (Cro)	Freq	6	41	152	0	1460
	per million	0.64	4.38	16.22	0	155
Croatian tourism corpus 3 (Cro)	Freq	24	23	117	0	668
	per million	4.06	3.89	19.81	0	113.09
Croatian tourism corpus 4 (Rovinj, Pula, Zagreb)	Freq	57	5	60	22	293
	per million	4.71	0.41	4.96	1.82	24.23
Corpus of Croatian tourism in Japanese (2nd method, URL)	Freq	3	1	6	2	145
	per million	23.21	7.74	46.42	15.47	1121.8
JaTenTen 2011, LUW sample	Freq	0	4	0	0	33
	per million	0	0,02	0	0	0.16

The results confirm the relevance of the target corpora specialized in the domain of tourism, as a majority of the specified terms appear in the corpora with much higher values than in the reference corpus, and some of them do not even appear in the reference corpus. The Corpus of Croatian tourism in Japanese (2nd method, URL) demonstrates the highest normalized values for all terms and confirms its relevance as a specialized corpus. Some of the targeted words are not expected to appear in some of the corpora (for example, corpus 1 targets the Istrian region, so the town of Osijek is not expected).

All of the corpora can provide relevant word sketch results, which are useful in discovering the most important collocations and patterns for a specific lexical item and could be well incorporated into the dictionary. Figure 8 shows word sketch results for the word 海 *umi* 'sea' with relevant collocations (e.g. 海の色 *umi no iro* 'the colour of the sea', クロアチアの海と山 *Kuroachia no umi to yama* 'Croatian sea and mountains') and links to examples in the corpus.

Finally, good dictionary examples could be retrieved from the data and were used for the Japanese-Croatian online learner's dictionary (Srdanović 2018). Figure 9 shows the dictionary entry ザグレブ風カツレツ *Zagurebu-fuu katsuretsu* 'zagrebački odrezak' with an example that used the Croatian tourism corpus 2 as a source.



海 Corpus of Croatian tourism in Japanese language freq = 137 (1,059.91 per million)

particle			を:verb			の:pronom			に:verb			pronom:の		
		86.13			25.55			24.09			15.33			12.41
を	26	8.48	眺める	6	11.54	色	5	12.03	浮かぶ	5	11.76	エメラルド	2	11.75
海を			海を眺めながら			海の色が			海に浮かぶ灯り			エメラルドグリーンの海		
から	9	8.48	離れる	3	11.26	幸	4	11.71	囲む	2	10.60	島周辺	1	10.82
海からの			一方、海を離れ内陸に入る			は手長エビと海の幸と山の			海に囲まれた			島周辺の海		
と	11	8.30	入る	3	11.06	上	5	11.04	向かう	2	10.54	遠浅	1	10.75
の海と			一方、海を離れ内陸に入ると空気は			海の上にカラフルな			海に向かう			遠浅の海		
だけでなく	1	7.97	臨む	2	10.79	景色	2	10.30	突出為る	1	10.54	ドゥブロヴニク周辺	1	10.75
海だけでなく			海を臨む			海の景色を望むレストラン			海に突出し			ドゥブロヴニク周辺の海		
の	37	7.74	一望出来る	2	10.75	アクティビティー	1	9.91	潜る	1	10.54	雲	1	10.68
海の			海を一望できるパレコニー			海のアクティビティシーズン			海に潜っ			雲の海		
に	21	7.53	渡る	2	10.57	海域	1	9.91	泳ぐ	1	10.47	オパティヤ	2	10.51
海に			海を渡っ			海の海域			海に泳ぎ			オパティヤの海		
が	7	6.50	楽しむ	3	9.86	犬	1	9.91	面する	2	10.44	前	1	10.47
の海が			海を楽しみ			海の犬			とても広く、海に面しているので晴天時			前の海		

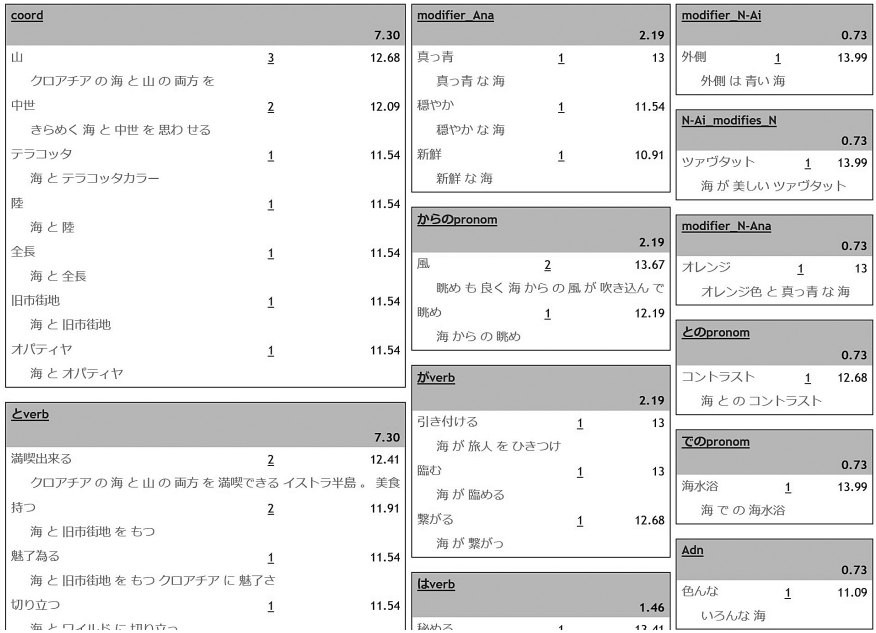


Figure 8: The word sketch results for the word 海 *umi* 'sea'

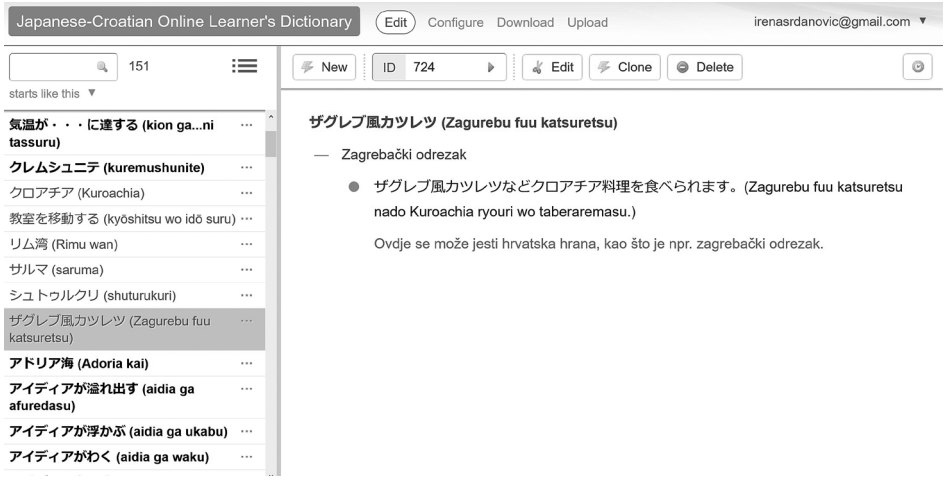


Figure 9: A dictionary entry with an example from the corpus

4. Conclusion and further work

This paper presented how specialized corpora of Croatian tourism in Japanese were created using two different methods. The corpora were used to extract keywords and terms, collocations, word sketches, and examples. As the results were well retrievable, they were used for sample dictionary entries for the Japanese-Croatian bilingual learner's dictionary specialized in tourism being created within the Lexonomy online dictionary platform.

The two methods have proven to be useful with some limitations to be further explored. The first method, using seed words to build a corpus, has resulted in a number of specialized corpora small in size but useful for extracting dictionary examples. As there is also noise in the data, there is a need to further clean up the data and explore the process for Japanese. The second method, collecting URL and building a specialized corpus, has proven to provide much cleaner data for the Japanese language and terms usable for dictionary were very well extracted; the corpus, however, could be further enlarged to be more usable for language description.

The research has indicated the need to improve foreign language and foreign culture-specific terms in language processing, in this case particularly referring to specific terms about Croatian tourism in Japanese, such as *Darumachia* 'Dalmatia (region)' being processed and segmented as Daruma 'Daruma, a Japanese traditional doll that symbolizes perseverance and good luck' + chiaa 'cheers'.

In future research, it is desirable to further analyze the results obtained using seed words and web pages, to try different methods, such as domains, one click, etc. and enlarge the specialized corpus and dictionary. The results of this research can be reused for other regions to cover tourism terminology in Japanese targeted at the Balkan and wider and create specialized corpora and dictionaries for Japanese language learners in combination with other languages.

References

Baroni Marko; Kilgarriff, Adam; Pomikalek, Jan; Rychly, Pave. 2006. WebBootCaT: a web tool for instant corpora. Eds. Corino, Elisa; Marello, Carla; Onesti, Cristina. *Euralex.* Edizioni dell'Orso. Torino. 123–131.

Baroni, Marko; Bernardini, Silvia. 2004. BootCaT: Bootstrapping corpora and terms from the Web. *Proceedings of LREC 2004*. Eds. Lino, Maria Teresa et al. European Language Resources Association. Lisbon. 1313–1316.

Baroni, Marko; Ueyama, Motoko. 2004. Retrieving Japanese specialized terms and corpora from the World Wide Web. *Proceedings of KONVENS*. Wien. 13–16.

Erjavec, Tomaž; Hmeljak Sangawa, Kristina; Srdanović, Irena. 2006. jaSlo, A Japanese-Slovene Learners' Dictionary: Methods for Dictionary Enhancement. Eds. Corino, Elisa; Marello, Carla; Onesti, Cristina. *Proceedings of the 12th EURALEX International Congress.* Edizioni dell'Orso. Turin.

Hmeljak Sangawa, Kristina; Erjavec, Tomaž. 2010. The Japanese-Slovene dictionary jaSlo: its developments, enhancement and use. *Studia Kognitiva* 10. 211–224. doi.org/10.11649/cs.2010.012.

Kilgariff, Adam. 2014. Finding terms in corpora for many languages with the Sketch Engine. *14th Conference of the European Chapter of the Association for Computational Linguistics*. Eds. Wintner, Shuly; Tadić, Marko; Babych, Bogdan. Association for Computational Linguistics. Gothenburg.

Kilgarriff, Adam. 2009a. Corpora in the classroom without scaring the students. *Proc. 18th Intntnl Symposium on English Teaching.* Taipei. www.kilgarriff.co.uk/Publications/2009-K-ETA-Taiwan-scaring.doc (accessed 16 March 2020).

Kilgarriff, Adam. 2009b. Simple maths for keywords. *Proceedings of Corpus Linguistics Conference CL2009.* Eds. Mahlberg, Michaela; González-Díaz, Victorina; Smith, Catherine. University of Liverpool. Liverpool.

Kilgarriff, Adam; Husák, Miloš; McAdam, Katy; Rundell, Michael; Rychlý, Pavel. 2008. GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of the 13th EURALEX International Congress*. Eds. Bernal, Elisenda; DeCesaris, Janet. Institut Universitari de Linguistica Aplicada – Universitat Pompeu Fabra. Barcelona. 425–432.

Kilgarriff, Adam; Rychly, Pavel; Smrž, Pavel; Tugwell, David. 2004. The Sketch Engine. *Euralex. Eds. Williams, Geoffrey; Vessier, Sandra*. Université de Bretagne-Sud – Faculté des lettres et des sciences humaines. Lorient. 105–116.

Srdanović, Irena. 2018a. Engaging Students in Creating Bilingual Online Learner's Dictionaries: General and Specializing in Tourism. *International Symposium Japanese Language Learning for New Generations: Book of Abstracts.* Eds. Srdanović, Irena; Matsuno, Naoyuki. Faculty of Humanities, Juraj Dobrila University of Pula. Pula.

Srdanović, Irena. 2018b. Towards Japanese Online Learner's Dictionary of Tourism. *International Symposium Tsukuba Day in Pula*. Faculty of Humanities, Juraj Dobrila University of Pula. Pula. (presentation).

Srdanović, Irena. 2019. Odabir prikladnih primjera za učenički rječnik: značajke GDEX-a za japanski jezik i mogućnosti uporabe. *Tabula: časopis Filozofskog fakulteta* 16. 175–197. doi.org/10.32728/tab.16.2019.14.

Srdanović, Irena; Erjavec, Tomaž; Kilgarriff, Adam. 2008. A web corpus and word sketches for Japanese. *Shizen gengo shori (Journal of Natural Language Processing)* 15/2. 137–159.

Srdanović, Irena; Kosem, Iztok. 2016. GDEX for Japanese: Automatic extraction of good dictionary examples. *GLOBALEX 2016 Lexicographic Resources for Human Language Technology*. Eds. Kernerman, Illan; Kosem, Iztok; Krek, Simon; Trap-Jensen, Lars. Portorož. 57–64.

Srdanović, Irena; Suchomel, Vít; Ogiso, Toshinobu; Kilgarriff, Adam. 2013. 百億語のコーパスを用いた日本語の語彙・文法情報のプロファイリング (Japanese Language Lexical and Grammatical Profiling Using the Web Corpus JpTenTen). *Proceeding of the 3rd Japanese corpus linguistics workshop*. NINJAL. Tokio. 229–238.

Ueyama, Motoko; Srdanović, Irena. 2018. *Digital Resources for Learning Japanese*. Bononia University Press. Bologna.

Yamasaki-Vukelić, Hiroshi. 2006. *Japansko-hrvatski, hrvatsko-japanski rječnik*. Dominović. Zagreb.

Internet pages

Statistics used in Sketch Engine. 2015. Lexical Computing Ltd. www.sketchengine.eu/documentation/statistics-used-in-sketch-engine/#KeyWords (accessed 16 March 2020).

Od specijaliziranih mrežnih korpusa do rječnika za neizvorne govornike

Sažetak

U ovom radu predstavljena su dva pristupa u stvaranju specijaliziranih mrežnih korpusa o hrvatskom turizmu na japanskome jeziku za potrebe izrade specijaliziranoga dvojezičnog japansko-hrvatskog rječnika za učenike japanskoga jezika. Oba pristupa koriste se tehnologijom WebBootCat unutar platforme Sketch Engine za automatsko stvaranje specijaliziranih mrežnih korpusa (Baroni i dr. 2006, Kilgarriff i dr. 2014). Prvi pristup stvara korpus iz odabranih riječi, odn. polaznih pojmova (eng. *seed words*) koje su najrelevantnije za temu. Drugi pristup skuplja brojne mrežne stranice sa sadržajima

o turizmu u određenim regijama, gradovima i mjestima u Hrvatskoj napisanim na japanskome jeziku, koje se zatim upotrebljavaju za stvaranje mrežnih korpusa. Oba pristupa pružaju specijalizirane internetske korpuse koji su, bez obzira na malu veličinu, korisni za profiliranje leksika u određenome području, u ovome slučaju u području turizma. U procesu stvaranja rječnika drugi se pristup pokazao osobito korisnim za odabir natuknica, dok su se oba pristupa pokazala vrlo korisnim za istraživanje i odabir autentičnih primjera iz korpusa. Uočeni su i određeni nedostatci u jezičnoj obradi podataka na japanskome jeziku, npr. pogreške u lematizaciji nekih pojmova i naziva svojstvenih određenoj kulturi te je ukazano na potrebu za usavršavanjem postojećih alata za jezičnu obradu podataka na japanskome jeziku. Japansko-hrvatski dvojezični rječnik namijenjen učenicima japanskoga jezika trenutačno je u eksperimentalnoj fazi te se učenici i nastavnici njime koriste i grade ga s pomoću otvorene platforme za izradu mrežnih rječnika Lexonomy (Mechura 2017). Osim što je rad na dvojezičnome rječniku koristan kao sredstvo za stručnu obuku u analizi i opisu jezika s pomoću suvremenih tehnologija (npr. korpusa, platforma za pretraživanje korpusa i izradu rječnika), važnost rječnika vidi se i u izobrazbi novih stručnjaka osposobljenih za rad u turizmu na japanskome jeziku, što je izrazito potrebno. U budućnosti bi se mogao primijeniti isti pristup za stvaranje specijaliziranih korpusa i rječnika za japanski i druge jezične parove.

Keywords: corpus building, BootCat technology, tourism domain, learners's dictionary, Sketch Engine, specialized web corpus of Croatian tourism in Japanese

Ključne riječi: izgradnja korpusa, tehnologija BootCat, područje turizma, rječnik za neizvorne govornike, Sketch Engine, specijalizirani mrežni korpus o hrvatskome turizmu na japanskome jeziku