http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2005 (Project)

Krajšavar — An Algorithm for Automatic Recognition of Abbreviations in English Dictionary Entries Compiled in a Dictionary of Abbreviations

Mojca Kompara Lukančič, *Faculty of Criminal Justice and Security and Faculty of Tourism, University of Maribor, Slovenia (mojca.kompara@gmail.com) (mojca.kompara@um.si) (https://orcid.org/0000-0003-2368-4161)*

Abstract: This article describes the use of Krajšavar — an algorithm for the automatic recognition of abbreviations in compiling English dictionary entries for a dictionary of abbreviations (*Slovar krajšav*), published in 2025 and financed by the Slovenian Research and Innovation Agency (ARIS). Together with the Slovenian dictionary of abbreviations (*Slovenski slovar krajšav*) published in 2023, the mentioned dictionary adopts a pioneering approach to the compilation of dictionaries in Slovenia as these are the first contemporary dictionaries of abbreviations. The *Slovar krajšav* was compiled based on an analysis of the characteristics of English dictionary entries for abbreviations, and according to the characteristics of the compilation methods used for bilingual dictionaries. The dictionary of abbreviations includes entries in 22 languages, with the most frequent being English, Italian and French. In this article the focus is on compiling English dictionary entries and using the Krajšavar algorithm.

Keywords: ABBREVIATIONS, ENGLISH, DICTIONARY, ALGORITHM, DICTIONARY OF ABBREVIATIONS, LANGUAGES, KRAJŠAVAR

Opsomming: Krajšavar — 'n algoritme vir die outomatiese herkenning van afkortings in Engelse woordeboekinskrywings wat in 'n afkortingswoordeboek saamgestel is. Hierdie artikel beskryf die gebruik van Krajšavar — 'n algoritme vir die outomatiese herkenning van afkortings in Engelse woordeboekinskrywings wat in 'n afkortingswoordeboek (*Slovar krajšav*) saamgestel is. Dit is in 2025 gepubliseer en gefinansier deur die Sloweense Agentskap vir Navorsing en Innovasie (ARIS). Saam met die Sloweense afkortingswoordeboek (*Slovenski slovar krajšav*) wat in 2023 gepubliseer is, verrig die woordeboek onder bespreking baanbrekerswerk in die samestelling van woordeboeke in Slowenië, aangesien hierdie publikasies die eerste eietydse afkortingswoordeboeke is. *Slovar krajšav* se samestelling berus op 'n ontleding van die eienskappe van Engelse woordeboekinskrywings vir afkortings, en is volgens die eienskappe van die samestellingsmetodes wat vir tweetalige woordeboeke gebruik word, uitgevoer. Die afkor-

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2005 (Project)

tingswoordeboek sluit inskrywings in 22 tale in, met Engels, Italiaans en Frans wat die algemeenste is. In hierdie artikel val die klem op die samestelling van Engelse woordeboekinskrywings en die gebruik van die Krajšavar-algoritme.

Sleutelwoorde: AFKORTINGS, ENGELS, WOORDEBOEK, ALGORITME, AFKORTINGS-WOORDEBOEK, TALE, KRAJŠAVAR

1. Introduction

In 2011, Kompara Lukančič and Holozan (2011) examined the automatic compilation of simple and complex dictionary entries in a dictionary of abbreviations. The research was based partly on the development of an algorithm for the automatic recognition of abbreviations in electronic texts (Kompara Lukančič 2010, 2011b). As Kompara Lukančič (2009, 2010, 2017, 2018) mentions, abbreviations — grouped, joined or unified compositions of letters — emerge rapidly and suddenly in a language, which makes it difficult to keep track of them and include them in general, bilingual, terminological, orthographic dictionaries (Kompara Lukančič 2009). This might be because some abbreviations remain in a language for only a limited period, such as *COVID* or *SARS*, while others are no longer recognised as abbreviations, e.g. *radar*. This may make it is difficult to determine which abbreviations should be included in a dictionary. Establishing a strategy or methodology for including abbreviations in dictionaries is also subject to discussions since the "life expectancy" of an abbreviation cannot be anticipated.

Kompara Lukančič (2009) argues that abbreviations are a growing phenomenon in the Slovenian language, with her research in a diachronic frame showing that abbreviations were included in orthographic dictionaries of the Slovenian language (Kompara Lukančič 2009, 2018). Kompara Lukančič (2018) emphasises the need to compile a separate dictionary of abbreviations for the Slovenian language, akin the abbreviations dictionaries already available in other languages, such as English, Italian, German and French.

The need to compile a Slovenian dictionary of abbreviations is further justified by the existence of such dictionaries in other languages, particularly English as it has the biggest number of published dictionaries of abbreviations (Kompara Lukančič 2009). Kompara Lukančič (ibid.) advocates for a contemporary dictionary of abbreviations in Slovenian that includes both Slovenian and foreign abbreviations, and whose structure and content is comparable to similar dictionaries in other languages, including English (cf. Paxton 1983, De Sola 1986, Fergusson 2000), French (cf. Faudouas 1990, Murith and Bocabeille 1992), German (cf. Koblischke 1983, Steinhauer 2005), Italian (cf. Malossini 1999, Righini 2001) and Spanish (cf. Galende 1997, 2001).

The aim of the paper is to expose the need for the compilation of a dictionary of abbreviations for Slovenian, specifically a dictionary in which both Slovenian and foreign abbreviations and expansions are provided, and that has dic-

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2005 (Project)

tionary entries composed of language qualifiers, filed qualifiers and translations. As an example of good practice in the compilation process of such a dictionary, similar dictionaries compiled for English, Italian and French is considered. In this paper, the focus is on the process of compiling the English dictionary entries in the *Slovar krajšav*, as English is the dominant foreign language among the 22 foreign languages included in the dictionary.

2. Dictionaries of abbreviations in the Slovenian language

In the past 25 years, abbreviations in the Slovenian language have been extensively discussed by Kompara Lukančič (2009, 2017, 2018, 2023a), Logar (2005), Verovnik and Logar (2006), Verovnik (2018) and Tonin (2022). The first dictionary of abbreviations *Kratice* was published in 1948 (Župančič 1948), followed by *Rečnik jugoslovenskih skraćenica* (Zidar 1971), and two more recent online dictionary attempts: *Slovarček krajšav* (Kompara Lukančič 2006) and *Slovar krajšav* (Kompara Lukančič 2011a). *Slovarček krajšav* (Kompara Lukančič 2006) was compiled as part of student work, with basic lexicographical knowledge yet without an appropriate structure. It is a collection of over 6 000 manually collected Slovenian and foreign abbreviations with Slovenian translations of foreign abbreviations. Among the foreign languages, are English, Italian, German, Spanish, Latin and French abbreviations, to name a few.

Figure 1: The dictionary entry for C (*Slovarček krajšav* 2006)

C	1. lat.: centum: rimska številka 100 2. Carboneum: kemijski simbol za ogljik 3. angl.: see: glej 4. stopinja Celzija 5. it.: codice: zakonik 6. programski jezik C

As evident in Figure 1 above, the dictionary entry is relatively simple. It is composed of an abbreviation, e.g. *C* in bold, followed by numbered expansions, e.g. (1), the abbreviated language qualifier (it.) for *Italian* next to (5), and the foreign expansions, e.g. *centum* and *Carboneum*, and the Slovenian translation given at (1) *rimska številka (Roman number)*. In the dictionary entry, both Slovenian and foreign expansions are given. Unfortunately, the dictionary does not have a concise structure, expansions should be alphabetically ordered, translations should be verified, and cross-references, encyclopaedic data and field qualifiers should be unified.

Slovar krajšav (Kompara Lukančič 2011a) is a dictionary of abbreviations that was compiled entirely automatically. Abbreviations and expansions were automatically extracted from a corpus containing Slovenian texts from the newspaper *Delo*, with this corpus containing five years' worth of *Delo* editions and a total of 60 million words. This automatic extraction was performed by using the algorithm for the automatic recognition of abbreviations and expansions in electronic texts (Kompara Lukančič 2009, 2010, 2017, 2018). After extraction,

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2005 (Project)

abbreviation–expansion pairs were manually verified, expansions were lemmatised, and language qualifiers were added automatically (Kompara Lukančič and Holozan 2011). The collection of 2 571 dictionary entries is available for free on the *Termania* website (https://www.termania.net/).

At the initial stage despite filtering Slovenian texts, Slovenian and foreign abbreviations were found among the dictionary entries. As shown in Figure 2 below, foreign abbreviations, such as English abbreviations and their expansions, are included in the dictionary. The foreign expansions are followed by the abbreviated language qualifier in brackets (en) *English*, which was added automatically (Kompara Lukančič and Holozan 2011). As seen from the dictionary entry, no Slovenian translations are provided, however. This information is crucial for dictionary users and would make the dictionary more usable and oriented towards them. By providing a translation the dictionary would solve users' translation problems. A similar approach is visible in the Italian dictionary of abbreviations (Righini 2001) but is not present in the German (Steinhauer 2005).

Figure 2: The dictionary entry for BA (*Slovar krajšav* 2011a)

BA
British Airways (en)
Bank Austria (en)
Budapest Airport (en)

Figure 3 shows that Slovenian abbreviations were also included in the collection. Slovenian expansions were lemmatised, and abbreviated language qualifier in brackets, such as (sl) *Slovenian*, were added automatically. As seen in Figure 3 below, no additional information or encyclopaedic data is included in the dictionary entry. The inclusion of such data could however be appreciated by dictionary users as it would give additional information concerning an abbreviation and expansion. It would clarify the meaning to the users, and for that reason such data should be seen as relevant and an improvement of the dictionary. A similar approach of inclusion of additional or encyclopaedic data is visible in the Italian dictionary of abbreviations (Righini 2001).

Figure 3: The dictionary entry for DA (*Slovar krajšav* 2011a)

DA
Demokratska akcija (sl)
Demokratična alternativa (sl)
državni aparat (sl)

Slovarček krajšav (Kompara Lukančič 2006) and *Slovar krajšav* (Kompara Lukančič

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2005 (Project)

2011a) are important to consider because they led to the preparation of the Slovenian dictionary of abbreviations *Slovenski slovar krajšav* (Kompara Lukančič 2023b). This new dictionary is a collection of 3 400 alphabetically ordered dictionary entries and over 4 000 expansions. The dictionary was compiled using the algorithm for the automatic recognition of abbreviations and expansions in Slovenian electronic texts as well as the algorithm for language detection and lemmatisation of Slovenian expansions. In the first and last phases of the compilation, abbreviations were collected manually.

The dictionary includes abbreviation–expansion pairs from the general language and several terminological fields. The dictionary was published in print in 2023 thanks to a grant given by the Slovenian Research and Innovation Agency (ARIS), and is also available online on the *Termania* website. As may be observed in Figure 4 below, the dictionary entry is relatively simple. The headword is followed by alphabetically ordered expansions and some additional information or explanations, such as *pevski glas* (singing voice), when needed. Filed qualifiers are not included. For expansions with two or more abbreviations, cross-references are included as a separate dictionary entry, for example *a*.

Figure 4: The dictionary entry for A (*Slovenski slovar krajšav* 2023b)

A adenin alt; pevski glas → a amper as; igralna karta masno število

3. The algorithm for the automatic recognition of abbreviations and expansions in electronic texts

Over the course of a decade, there has been a shift from a dictionary compilation process where an automated process was used to some degree, to the current advance towards a fully automatised process (Rundell 2023). Rundell (2023) describes some semi-automated projects, the gradual progress towards new approaches in compiling dictionaries (as in post-editing lexicography), the role of lexicographers in post-editing (namely evaluating and refining the first dictionary draft which was automatically generated and imported into a dictionary writing and editing system), and finally the arrival of algorithms in dictionary compilation.

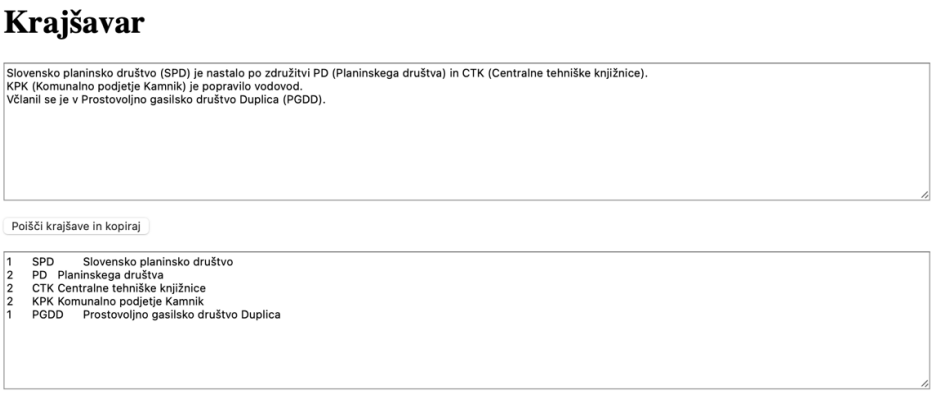
Algorithms for the automatic recognition of abbreviations and expansions in electronic texts have been researched for nearly three decades. The first such algorithm was developed in 1999 by Taghva and Gilbreth (1999). This pioneering approach was followed by Yeates (1999), Larkey et al. (2000), Park and Byrd (2001), Schwartz and Hearst (2003), Zahariev (2004), Xu and Huang (2005), Zhou, Torvik

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2005 (Project)

and Smalheiser (2006), Šatev and Nikolov (2008), and Kompara Lukančič (2009, 2018). These algorithms focused on recognising abbreviations and expansions in English texts, mainly in the medical field (Schwartz and Hearst 2003).

Kompara Lukančič (2011b, 2018) developed the first algorithm for recognising abbreviations and expansions in Slovenian. This algorithm has since been modified, improved and adapted several times (Kompara Lukančič 2009, 2011b, 2018), also for filtering English texts. With the support of Peter Holozan, an IT specialist who developed the latest version of the algorithm's user interface shown in Figure 5 (below), the Krajšavar algorithm is currently capable of filtering and extracting abbreviations and expansions from both Slovenian and English texts.

Figure 5: Krajšavar — an algorithm for the automatic recognition of abbreviations and expansions in electronic texts



The final preparation of the algorithm entailed several steps. It started with establishing basic rules for recognising abbreviations that were based on abbreviations' characteristics in both Slovenian and English. This was followed by setting basic rules for recognising an expansion, and finally implementing all the concepts in a user-friendly digital form.

The Krajšavar algorithm was developed, initially for the author's purposes, in the interface shown above in Figure 5. The user interface is composed of two windows: in the first one, the text is included for filtration, while in the second one, the abbreviations and expansions appear. Behind this interface, a complicated algorithm operates that first cuts the text into separate words, treating punctuation marks as separate words. The algorithm then filters the words from the text in search of open brackets. Upon finding open brackets, the algorithm continues to search for the first closed bracket that follows the first open bracket. If between the brackets there is only one word that starts with an uppercase letter (or is entirely uppercase) and is made up of at least two letters, the algorithm assumes it is an abbreviation.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2005 (Project)

To find expansions, the algorithm looks for words preceding the brackets and tries up to 10 words placed them. It then searches for the first word that matches the first letter of the abbreviation placed after open brackets (in uppercase). However, this procedure does not work well in languages like English in which words such as titles and proper names start with an uppercase letter. The algorithm tries to find abbreviations by following the same rule, namely if there are at least two words within brackets and the first letter of the first word matches the first word before the brackets, it assumes the abbreviation is before the brackets and the abbreviation's expansion is within the brackets.

The Krajšavar algorithm does not need an external server, is written in JavaScript, and included in the file krajšave.html. By clicking "Poišči krajšave in kopiraj" (find abbreviations and copy), the function "isciKrajsave" (find abbreviations) is recalled. In such a way, the text is read from the input file and sent to the function of tokenisation that cuts the text into words with the help of the regular expression. In the implementation in Python, an external library nltk.tokenize was used, in JavaScript the tokenisation was written manually to avoid any unnecessary external dependencies that might complicate the use of the webpage elsewhere. The candidates for abbreviations are given in the sequence "result" that at the end is written in the output file. At the very end, the sequence is copied onto the clipboard to enable further work and editing.

4. The compilation for the dictionary of abbreviations *Slovar krajšav*

Over the past 20 years, the need for a dictionary of abbreviations to be compiled has been discussed extensively (Kompara Lukančič 2009, 2010, 2011b, 2017, 2018; Kompara Lukančič and Holozan 2011). The compilation of the Slovenian dictionary of abbreviations, *Slovenski slovar krajšav* (Kompara Lukančič 2023b), led to the compilation of a more comprehensive work, namely a dictionary of abbreviations consisting of foreign abbreviations from 22 languages that have been compiled in a single volume. This dictionary provides the expansions of these abbreviations in the foreign languages along with Slovenian translations and, if they exist, Slovenian abbreviation equivalents. Preparation of the dictionary was made possible with the financial support of the ARIS, and the dictionary was published in 2025.

On a macrostructural level, the dictionary includes approximately 3 500 alphabetically ordered dictionary entries and over 4 200 expansions, noting that some abbreviations can have several expansions. The dictionary structure is based on the structure used in the Slovenian dictionary of abbreviations *Slovenski slovar krajšav* (Kompara Lukančič 2023b). Work on compiling the dictionary spanned over two decades. The dictionary was initially compiled using the algorithm for the automatic recognition of abbreviations and expansions in electronic texts (Kompara Lukančič 2009, 2010), and also the algorithm for language recognition and lemmatisation, with the final stage of the compilation process being completed manually. In later years, the algorithm was adapted

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2005 (Project)

for filtering English texts, and the material obtained from these texts was added to the dictionary.

In the process of compiling English dictionary entries for the *Slovar krajšav*, the characteristics of dictionary entries in English dictionaries of abbreviations were examined (Kompara Lukančič 2023a). An extensive analysis of English dictionaries of abbreviations published between 1942 and 2019 was conducted in 2009 and updated in 2023 (Kompara Lukančič 2009, 2023a). The Krajšavar algorithm was introduced in the final phase of the compilation process. As discussed in Kompara Lukančič (2023a), while compiling the English entries for the *Slovar krajšav* based on the examination of dictionaries of abbreviations, the alphabetical order of expansions found in *Everyman's Dictionary of Abbreviations* (Paxton 1983) was retained, as well as language qualifier *angl. English*. The numerical order preceding an expansion was not preserved, however. Similar to *A Dictionary of Abbreviations: With Special Attention to War-Time Abbreviations* (Partridge 1942), some additional information or encyclopaedic data were added to a dictionary entry. As in the *Everyman's Dictionary of Abbreviations* (Paxton 1983), field qualifiers, abbreviated and provided in brackets were also used in a dictionary entry, for example (voj.) *army*. Cross-references and expansions with more than one abbreviation were also included as separate dictionary entries, which is similar to the approach in the *Everyman's Dictionary of Abbreviations* (Paxton 1983). Expansions of one abbreviation are written within a single dictionary entry, which differs from the *Dictionary of Abbreviations in Medical Sciences* (Heister 1989) where every expansion has a separate dictionary entry. English expansions are also followed by translations into Slovenian, like in *The Barnhart Abbreviations Dictionary* (Barnhart 1995), *The New Penguin Dictionary of Abbreviations* (Fergusson 2000), and the *Dictionary of Financial Abbreviations* (Paxton 2003), where foreign expansions are translated.

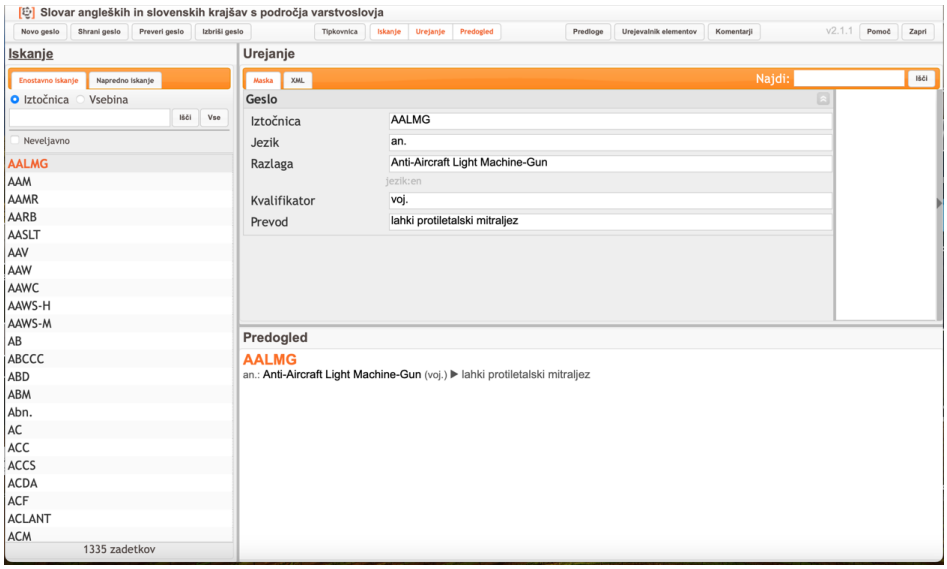
5. The use of the Krajšavar algorithm in compiling the *Slovar krajšav*

As mentioned, the Krajšavar algorithm was introduced in the final stage of the compilation process of the *Slovar krajšav*. The research was based on a collection of English texts following a text typology of texts from the field of criminal justice and security. These texts were collected using Google and Google Scholar as search engines and in line with a text typology for criminal justice and security (Kompara Lukančič, forthcoming). A total of 210 texts were filtered using the Krajšavar algorithm, resulting in 1 335 abbreviation–expansion pairs. The text filtration was followed by the automatic transfer of all abbreviation–expansion pairs into the *Termania* dictionary editing interface, as shown in Figure 6 below. As evident in Figure 6, the obtained abbreviations and expansions were included in the *Termania* dictionary mask, and language qualifiers, such as *an. English*, as well as field qualifiers, such as *(voj.) army*, were added together with the Slovenian translations. The database was named the Dictionary of English and Slovenian Abbreviations from the Field of Criminal Justice and Security

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2005 (Project)

(*Slovar angleških in slovenskih krajšav s področja varstvoslovja*). Each dictionary entry in the preview section consists of a headword, namely the abbreviation (for example, *AALMG*), followed by a language qualifier that is abbreviated (such as *an.*), the foreign expansion (e.g., *Anti-Aircraft Light Machine-Gun*), a field qualifier, which is abbreviated in brackets, such as *(voj.)*, and finally the symbol ▶ denoting the Slovenian translation.

Figure 6: Example of a dictionary entry for AALMG



If an abbreviation has several expansions, all of them are included in the same dictionary entry, with the alphabetical order preserved. Symbols used in the process of compiling the *Slovar krajšav* are explained in the paragraph below. Abbreviation–expansion pairs obtained with Krajšavar were manually verified before being included in the *Slovar krajšav*. The structure of the dictionary entries in the *Slovar krajšav* is explained in the following section.

6. Examples of English dictionary entries in the *Slovar krajšav*

Figure 7 below presents a dictionary entry showing that the abbreviation *CC* has seven alphabetically ordered expansions in the dictionary entry. The dictionary entry consists of an entry word, followed by an abbreviated language qualifier (such as *angl.* – *English*), the English (or other languages) expansion, and an abbreviated field qualifier (such as *(ekon.) economics*) in brackets.

Some special symbols are used in the dictionary entry. Expansions are

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2005 (Project)

followed by the symbol ►, denoting a translation, and the symbol ◊, indicating a descriptive sentence. The symbol ◊ is used for culturally specific expansions that do not have a translation equivalent, and thus a descriptive sentence is used instead. Examples of this are *član kanadskega viteškega reda* (Member of the Order of Knights of Canada) and *polje v e-poštnemu sporočilu za naslovnike, ki prejmejo kopijo pošte* (a field in the email message for other recipients who receive a copy of the email). The symbol ○ is used when a Slovenian equivalent abbreviation of the English one exists, for example *Kp, OZS* and *POVC*. In these cases, cross-references are also provided and included as separate dictionary entries. If an expansion has multiple abbreviations, such as *CC, c. c. and cc* for *carbon copy*, they are introduced in the dictionary entry with the word *tudi* (also) and included as separate dictionary entries for each abbreviation.

Figure 7: Example of the dictionary entry for CC

CC
angl.: carbon copy; tudi c. c., cc ◊ polje v e-poštnemu sporočilu za naslovnike, ki prejmejo kopijo pošte ○ Kp
angl.: Chamber of Craft (ekon.) ► Obrtna zbornica Slovenije ○ OZS
angl.: combat command (voj.) ► bojno poveljstvo
angl.: command center (voj.) ► poveljniški center ○ POVC
angl.: Companion of the Order of Canada ◊ član kanadskega viteškega reda
fr.: corps consulaire ► konzularni zbor
it.: Carabinieri (polic.) ► karabinjerji

As shown in Figure 8 below, the official translations, such as *srednjeevropski čas* (Central European Time), are mainly provided for English expansions that have previously been translated and can be found in dictionaries, databases and so forth. Where an official translation of an English expansion could not be found, a descriptive sentence containing additional information was used, for example *skupna carinska tarifa za uvoz iz držav nečlanic EU* (common external tariff).

Figure 8: Example of the dictionary entry for CET

CET
angl.: Central European Time (geo.) ► srednjeevropski čas
angl.: common external tariff (ekon.) ◊ skupna carinska tarifa za uvoz iz držav nečlanic EU

7. Discussion

As outlined in the introduction, the structure of the *Slovar krajšav* is based on the earlier compiled *Slovenski slovar krajšav* (Kompara Lukančič 2023b) and an

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2005 (Project)

analysis of the characteristics of dictionaries of abbreviations in English (Kompara Lukančič 2023a). As seen in the above examples, in the dictionary entry the alphabetical order is preserved, like with the *Everyman's Dictionary of Abbreviations* (Paxton 1983). Additionally, the structure of the English dictionary entry is relatively simple and is composed of a headword, followed by an abbreviated language qualifier, an English expansion, a field qualifier in brackets (if the abbreviation is not generic, such as in the case of *carbon copy*), and a translation, equivalent or descriptive sentence in Slovenian, as well as a Slovenian abbreviation, if one exists, such as *Kp, OZS* and *POVC*.

All elements are introduced in the dictionary entry with the use of special symbols, as explained in the previous section. In the printed version of the dictionary, the use of special symbols as well as the dictionary's macro and microstructure are explained in the introduction. There dictionary users will find examples of dictionary entries and an explanation of the structure of elements of an entry.

In some entries, additional or encyclopaedic data are included in line with a specific terminological dictionary of abbreviations, namely *A Dictionary of Abbreviations: With Special Attention to War-Time Abbreviations* (Partridge 1942). Abbreviated field qualifiers in brackets, cross-references, and expansions with multiple abbreviations are included in the same manner as in *Everyman's Dictionary of Abbreviations* (Paxton 1983), while Slovenian translations of the expansions are included in the dictionary entries, similar to the approach in *The Barnhart Abbreviations Dictionary* (Barnhart 1995). English abbreviations represent the most frequent foreign abbreviations collected in the *Slovar krajšav* (cf. Figures 7 and 8).

By including abbreviation–expansion pairs obtained with the Krajšavar algorithm (cf. Figure 6), a larger number of abbreviations were included in the dictionary, namely those from the field of criminal justice and security. Filtering texts from this field, gathered in line with the text typology classification (Kompara Lukančič, forthcoming), ensured a broader range of abbreviations from the field, as well as the inclusion of up-to-date abbreviations from the field since the filtered texts came from a more recent time frame, namely the last five years. This approach of including abbreviations by using the Krajšavar algorithm also lowered the possibility of forgetting to include relatively known and new abbreviations that might have been overlooked in a manual compilation. In a way, the algorithm's application ensures the outcome is up-to-date and not missing relevant abbreviations. Moreover, thanks to the text typology classification for criminal justice and security (Kompara Lukančič, forthcoming), a similar classification of texts can be applied to other language for special purposes fields, such as tourism, medicine and economics, enabling the inclusion of abbreviations from other fields as well.

8. Conclusion

The article describes the application of the Krajšavar algorithm in the process of

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2005 (Project)

compiling the *Slovar krajšav* and shows the need for a dictionary of abbreviations to be compiled in the Slovenian language. Dictionaries of abbreviations for the Slovenian language are presented in a synchronic and diachronic framework (cf. Kompara Lukančič 2018), namely two outdated dictionaries *Kratice* (Župančič 1948) and *Rečnik jugoslovenskih skraćenica* (Zidar 1971), and three more recent online dictionary attempts, namely *Slovarček krajšav* (Kompara Lukančič 2006), *Slovar krajšav* (Kompara Lukančič 2011a), and the most recently published *Slovenski slovar krajšav* (Kompara Lukančič 2023b).

The publication of the *Slovenski slovar krajšav* (Kompara Lukančič 2023b) led to the compilation of the single-volume dictionary of abbreviations *Slovar krajšav*, which is a collection of 3 500 alphabetically ordered dictionary entries and over 4 200 expansions from 22 foreign languages. The article outlines the overall compilation of the *Slovar krajšav* and discusses examples of dictionary entries for English abbreviations. As shown by the presented examples, a dictionary entry is composed following the compilation process used in previously published dictionaries the *Slovarček krajšav* (Kompara Lukančič 2006), *Slovar krajšav* (Kompara Lukančič 2011a) and *Slovenski slovar krajšav* (Kompara Lukančič 2023b), coupled with the characteristics of a range of English dictionaries of abbreviations (Kompara Lukančič 2009, 2018).

The compilation process took almost two decades to complete and included the application of several algorithms for lemmatisation, language detection and the automatic recognition of abbreviations. In the final preparation steps, the dictionary was compiled manually and with the help of Krajšavar algorithm, which allowed for the inclusion of abbreviations from a specialised field, as well as relevant abbreviations obtained from a range of texts following the relevant text typology. The *Slovar krajšav,* together with the *Slovenski slovar krajšav* (Kompara Lukančič 2023b), therefore represents an important contribution to the linguistic framework of abbreviations for the Slovenian language.

References

Barnhart, K.R. (Ed.). 1995. *The Barnhart Abbreviations Dictionary.* New York: John Wiley & Sons.

De Sola, R. 1986. *Abbreviations Dictionary.* New York: Elsevier.

Fergusson, R. 2000. *The New Penguin Dictionary of Abbreviations.* London: Penguin.

Faudouas, J.-C. 1990. *Dictionnaire des abréviations courantes de la langue française*. Paris: La Maison du Dictionnaire.

Galende, J.C. 1997/2001. *Diccionario general de abreviaturas españolas*. Madrid: Editorial Verbum.

Heister, R. 1989. *Dictionary of Abbreviations in Medical Sciences.* Berlin/Heidelberg/New York: Springer.

Koblischke, H. 1983. *Großes Abkürzungsbuch*. Third edition. Leipzig: VEB Bibliographisches Institut.

Kompara Lukančič, M. 2006. *Slovarček krajšav*. Ljubljana: Inštitut za slovenski jezik Fran Ramovš ZRC SAZU.

Kompara Lukančič, M. 2009. Prepoznavanje krajšav v besedilih. Weiss, P. (Ed.). 2009. *Jezikoslovni zapiski* 15(1–2): 95-112. Ljubljana: Inštitut za slovenski jezik Frana Ramovša.

Kompara Lukančič, M. 2010. Krajšavni slovarji. Weiss, P. (Ed.). 2010. *Jezikoslovni zapiski* 16(2): 111-129. Ljubljana: Inštitut za slovenski jezik Frana Ramovša.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2005 (Project)

Kompara Lukančič, M. 2011a. *Slovar krajšav*. Kamnik: Amebis, Termania.

Kompara Lukančič, M. 2011b. Razvoj algoritma za samodejno prepoznavanje krajšav in krajšavnih razvezav v elektronskih besedilih. Weiss, P. (Ed.). 2011. *Jezikoslovni zapiski: Zbornik Inštituta za slovenski jezik Frana Ramovša* 17(2): 107-122. Ljubljana: Inštitut za slovenski jezik Frana Ramovša.

Kompara Lukančič, M. 2017. Zasnova novega slovarja krajšav. Weiss, P. (Ed.). 2017. *Jezikoslovni zapiski* 23(1): 77-92. Ljubljana: Inštitut za slovenski jezik Frana Ramovša.

Kompara Lukančič, M. 2018. *Sinhrono-diahroni pregled krajšav v slovenskem prostoru in sestava slovarja krajšav*. Maribor: Univerzitetna založba Univerze.

Kompara Lukančič, M. 2023a. Compilation of English Entries in the *Contemporary Slovene Dictionary of Abbreviations*. *International Journal of Lexicography* 36(2): 195-210. https://doi.org/10.1093/ijl/ecac016

Kompara Lukančič, M. 2023b. *Slovenski slovar krajšav*. Maribor: Univerza v Mariboru, Univerzitetna založba.

Kompara Lukančič, M. 2025. *Slovar krajšav*. Maribor: Univerza v Mariboru, Univerzitetna založba. (in press). https://press.um.si/index.php/ump/catalog/book/948

Kompara Lukančič, M. Krajšavar — An Algorithm for Recognizing English Abbreviations in Texts Related to Criminal Justice and Security (forthcoming).

Kompara Lukančič, M. and P. Holozan. 2011. What is Needed for Automatic Production of Simple and Complex Dictionary Entries in the First Slovene Online Dictionary of Abbreviations Using Termania Website. Kosem, Iztok and Karmen Kosem (Eds.). 2011. *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex 2011, Bled, Slovenia, 10–12 November 2011*: 140-146. Ljubljana: Trojina, Institute for Applied Slovene Studies.

Larkey, L.S., P. Ogilvie, M.A. Price and B. Tamilio. 2000. Acrophile: An Automated Acronym Extractor and Server. *Proceedings of the Fifth ACM Conference on Digital Libraries, San Antonio, Texas, USA, 2–7 June 2000:* 205-214. New York: Association for Computing Machinery.

Logar, N. 2005. Norma v slovarju sodobne slovenščine: zloženke in kratice. Družboslovne razprave (Ed.). 2005. *Družboslovne razprave* 21(48): 211-225. Ljubljana: Slovensko sociološko društvo: Fakulteta za družbene vede.

Malossini, A. (Ed.). 1999. *Dizionario delle sigle e degli acronimi*. Milan: A. Vallardi.

Murith, J. and J.-M. Bocabeille. 1992. *Dictionnaire des abréviations et acronymes*. Second edition. Paris: Technique et documentation — Lavoisier.

Park, Y. and R.J. Byrd. 2001. Hybrid Text Mining for Finding Abbreviations and Their Definitions. *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, Pittsburgh, Pennsylvania, 3–4 June 2001*. https://aclanthology.org/W01-0516.pdf

Partridge, E. 1942. *A Dictionary of Abbreviations: With Special Attention to War-Time Abbreviations.* London: Routledge.

Paxton, J. 1983. *Everyman's Dictionary of Abbreviations.* London: Dent & Sons.

Paxton, J. 2003. *Dictionary of Financial Abbreviations.* New York: Routledge.

Righini, E. 2001. *Dizionario di sigle, abbreviazioni e simboli*. Bologna: Zanichelli.

Rundell, Michael. 2023. Automating the Creation of Dictionaries: Are We Nearly There? *ASIALEX 2023: Lexicography, Artificial Intelligence, and Dictionary Users (Asialex 2023), 22–24 June 2023, Seoul, Korea:* 9-17. Seoul: Yonsei University.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2005 (Project)

Schwartz, A.S. and M.A. Hearst. 2003. A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Texts. *Proceedings of the Pacific Symposium on Biocomputing* 8: 451-462. http://psb.stanford.edu/psb-online/proceedings/psb03/schwartz.pdf

Steinhauer, A. 2005. *Das Wörterbuch der Abkürzungen*. Mannheim: Duden Verlag.

Šatev, V. and N. Nikolov. 2008. Using the Web as a Corpus for Extracting Abbreviations in the Serbian Language. Erjavec, T. and J. Žganec Gros (Eds.). 2008. *Jezikovne tehnologije: Zbornik 11. mednarodne multikonference Informacijska družba — IS 2008*: 75-79. Ljubljana: Institut Jožef Stefan.

Taghva, K. and J. Gilbreth. 1999. Recognizing Acronyms and Their Definitions. *International Journal on Document Analysis and Recognition* 1(4): 191-198. https://doi.org/10.1007/s100320050018

Tonin, G. 2022. Kratični termini in priporočila za njihovo uporabo. *Jezik in Slovstvo* 67(1–2): 209-221. https://doi.org/10.4312/jis.67.1-2.209-221

Verovnik, T. 2018. Obravnava kratic v prenovljenih pravopisnih pravilih. *Jezikoslovni zapiski* 24(2): 43-54. https://doi.org/10.3986/jz.v24i2.7104

Verovnik, T. and N. Logar. 2006. O jeziku, stilu i utjecaju slovenskih tiskanih oglasa. Granić, J. (Ed.). 2006. *Jezik i mediji: jedan jezik: više svjetova:* 743-752. Zagreb, Split: Hrvatsko društvo za primijenjenu lingvistiku.

Xu, J. and Y. Huang. 2005. A Machine Learning Approach to Recognizing Acronyms and Their Expansions. *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, Guangzhou, China, 18–21 August 2005. Volume 4: 2313-2319. https://doi.org/10.1109/ICMLC.2005.1527330

Yeates, S. 1999. Automatic Extraction of Acronyms from Text. *Proceedings of the Third New Zealand Computer Science Research Students' Conference, Te Kohinga Marama Marae, Hamilton, New Zealand, 6–9 April, 1999. University of Waikato 1999:* 117-124.

Zahariev, M. 2004. *A (Acronyms).* Unpublished PhD Thesis. School of Computing Science, Simon Fraser University.

Zhou, W., V.I. Torvik and N.R. Smalheiser. 2006. ADAM: Another Database of Abbreviations in MEDLINE. *Bioinformatics* 22*:* 2813-2818. https://doi.org/10.1093/bioinformatics/btl480

Zidar, J. 1971. *Rečnik jugoslovenskih skraćenica.* Beograd: Međunarodna politika.

Župančič, J. 1948. *Kratice.* Ljubljana: Državna založba Slovenije.