Dictionary examples

Iztok Kosem

Abstract

In this paper, the role of examples in dictionary entries is presented, and an overview provided of relevant studies into the use and usefulness of examples. We put forward the different ways of presenting examples in general monolingual dictionaries, list the characteristics of a good dictionary example, and discuss the different methods of finding good examples. The focus then turns to the role and characteristics of examples in the proposal for a dictionary of modern Slovene, the methods for their extraction, and the procedures to be followed for saving examples to the dictionary database and archiving them, before concluding with the different visualisation options for the (online) dictionary.

Keywords: dictionary examples, good examples, automatic extraction, visualisation, dictionary database

1 INTRODUCTION

Examples are one of the most important parts of a dictionary entry, as they are used for exemplifying the use of words, collocations, compounds, phraseology and so on in context, i.e. in real language. Putting the words back into context is vital for a dictionary, since the majority of dictionary content is decontextualized.

The paper first describes the role of examples in a dictionary, and makes an overview of research into the usefulness of dictionary examples. This is followed by the presentation of different ways of example presentation in monolingual dictionaries of Slovene and other languages. Next, the characteristics of good dictionary examples are presented and different methods for finding them are described. The paper focuses on the role and characteristics of examples in the proposed *Dictionary of Modern Slovene Language* (DMSL), their acquisition from corpora and ways of recording them in the dictionary database. Visualization of examples in the dictionary is also briefly discussed. The conclusion summarizes the main points of the paper and considers the future role of examples in dictionaries and related resources.

2 THE ROLE OF EXAMPLES IN A DICTIONARY

The role of examples concerns two aspects of dictionary use: receptive and productive. The receptive aspect, which examples are primarily intended for, is to supplement definitions, which is why examples first and foremost need to contain information related to the meaning they attest. As argued by Atkins and Rundell (2008: 454), it is sometimes difficult for the user to understand the definition without reading the examples. Examples can also be useful when navigating through (long) entries, as the users can "identify the particular sense they are seeking by finding examples that are similar to the one they need or have in front of them" (Fox 1987: 137).

The productive role of dictionary examples is to attest the syntactic patterns, valency, collocations and other characteristics of the headword (Húmble 2001), which are supposed to help the users when writing or, less often, speaking. Examples intended for production are found mainly in dictionaries for L2 learners, e.g. advanced learners' dictionaries or dictionaries for younger native speakers, such as school dictionaries.

Studies into dictionary examples have mainly focused on their value for language production of non-native speakers. The most commonly used research method involves asking the subjects to use (unknown) words in a sentence, and consulting dictionaries or selected dictionary entries in the process. The subjects are grouped into those that are provided only with definitions, and those that are provided with definitions and examples; some studies (e.g. Frankenberg-Garcia 2012; 2014) also

include a group of subjects that are provided only with examples. The findings of the majority of studies (Summers 1988; Laufer 1993; Nesi 1996; Al-Ajmi 2008) are not very encouraging, as they show that examples do not have considerable added value for the encoding needs of the users. However, as Frankenberg-Garcia (2012) pointed out, the aforementioned studies have two key methodological shortcomings: firstly, despite studying the productive value of examples, the studies contain tasks in which the subjects need to first decipher the meaning of an unknown word and then use that word in a sentence. This means that the tasks include both receptive and productive dictionary use, which is a rare form of dictionary use. Secondly, using unknown words for testing productive use does not reflect actual dictionary use and language production in general, as people rarely use completely new words when writing (Laufer 1993: 138),

Frankenberg-Garcia (2012; 2014) improved the methodology of previous studies by clearly distinguishing between testing the receptive and productive roles of dictionary examples, and also by using examples for reception and examples for production, respectively. The subjects were divided into four groups: the control group (without a dictionary), the group that was provided only with definitions, the group that was provided with one corpus example, and the group that was provided with several corpus examples. Her findings were that several corpus examples are almost equally valuable as the definition when trying to understand the meaning of a word, and that for encoding use several corpus examples are much more useful than one example, while in general examples are much more useful than definitions.

There are very few studies that research how frequently the users consult examples. In a study that involved his students, Béjoint (1981) found that they consulted examples quite frequently. Similar are findings of Kosem's study among 620 students (449 native speakers and 171 non-native speakers of English) at Aston University; examples were the fourth most frequently consulted part of the dictionary entry (after definitions, pronunciation and synonyms), and, when considering only non-native speakers, examples were the second most frequently consulted part of the entry (after the definitions).

3 EXAMPLES IN GENERAL MONOLINGUAL DICTIONARIES

An analysis of the treatment and form of examples in general monolingual dictionaries[1] shows three different groups of dictionaries. The first includes those that offer

1 The analysis included only online dictionaries. The dictionaries can have paper versions or were originally published in the paper format, but the list also includes dictionaries that exist only online (e.g. the *Comprehensive Dictionary of Polish* and *Comprehensive Dictionary of Dutch*). A detailed analysis of the treatment of examples in dictionaries of Slovene is provided after the description of all three groups of dictionaries.

examples mainly in the form of partial sentences (or phrases) and occasionally also as whole sentences (e.g. the Spanish monolingual dictionary). Some dictionaries limit the use of examples only to certain (sub)senses or phrases. The information on the source of the example is rarely provided (there are exceptions, such as the *Explanatory Dictionary of Estonian*). Such treatment of examples is often found in dictionaries that are not corpus-based, and were originally conceived for print and later transferred to the online format. A few recently published dictionaries have also adopted such treatment, mainly those that were conceptualised according to the lexicographic approaches of the 20th century. The group includes dictionaries such as the *Dictionary of Literary Czech*[2] (DLC; Slovník spisovného jazyka českého, 1989), *Royal Spanish Academy Dictionary of Spanish*[3] (RSADS; Diccionario de la lengua Española de la Real Academia Española, 2014), *Explanatory Dictionary of Estonian*[4] (EDE; Eesti keele seletav sõnaraamat, 2007) and the *Croatian Encyclopaedic Dictionary*[5] (CED; Hrvatski enciklopedijski rječnik, 2003).

In the second group are dictionaries that offer mainly whole-sentence (corpus) examples, examples in the form of partial sentences are rare or not used at all. This treatment of examples can be found in English dictionaries published by Oxford (Oxford Dictionaries;[6] ODE), Macmillan (*Macmillan English Dictionary*; MED[7]) and Merriam-Webster (*The Merriam-Webster Online Dictionary*; MWOD[8]), the *Dictionary of Contemporary Danish* (Den Danske Ordbog;[9] DDO), the *Comprehensive Dictionary of Polish*[10] (CDP; Wielki Słownik języka Polskiego) and the *Comprehensive Dictionary of Dutch*[11] (CDD; Algemeen Nederlands Woordenboek). Nonetheless, dictionaries differ in the manner they present the examples. In MED and DCD, whole-sentence examples are presented within the entry, under each sense, subsense, phrase and so on. MWOD and ODE use both whole-sentence examples and excerpts, but clearly distinguish between them in terms of their presentation in the entry. Excerpts are offered under senses and subsenses at the first level, so immediately upon opening the entry, whereas whole-sentence examples (for all senses) are provided together at the end of the entry (MWOD) or available under senses by clicking on "More example sentences" (ODE). A somewhat less prominent role is given to examples by CDP and CDD, where these are not shown upon opening the entry and are only available on a click (CDD) or in a separate tab (CDP). These two dictionaries also provide the information on the source of the example.

2 http://ssjc.ujc.cas.cz (the online version available since 2011).
3 http://lema.rae.es/drae
4 http://en.eki.ee/dict/ekss
5 Accessible through the Croatian Dictionary Portal http://hjp.novi-liber.hr.
6 http://www.oxforddictionaries.com/
7 http://www.macmillandictionary.com/
8 http://www.merriam-webster.com
9 http://ordnet.dk/ddo
10 http://wsjp.pl
11 http://anw.inl.nl

The third group includes portals such as German DWDS[12] (Das Digitale Wörterbuch der deutschen Sprache) that offer on one page the information from dictionaries, corpora and other relevant resources.[13] The most important characteristic of this group is the link between dictionaries and corpora, with corpora being a source of an abundant number of examples, especially considering Frankenberg-Garcia's findings about the benefits of multiple examples for dictionary users. A shortcoming of such portals is in the large amount of information they provide, which often makes it difficult for the users to interpret and correctly use them.[14]

As far as dictionaries of Slovene are concerned, the *Dictionary of Slovene Literary Language* (DSLL) and its successor DSLL2 belong to the first group of dictionaries, offering examples as excerpts. The excerpts were taken from texts or were in some cases invented.[15] At least for DSLL, this finding is not surprising, given that the dictionary was made before the corpus lexicography era. However, DSLL contains a considerable quantity of examples, much more than comparable dictionaries of other languages, including recently published ones (e.g. RSADS). Examples were one of the most heavily affected parts of dictionary entries during the preparation of DSLL2, as the examples from DSLL were modified or replaced due to social changes, or completely new examples were added. As noted by Krek (2014: 146), however, changes in existing examples are often not appropriate or necessary, or completely new examples do not bring any added-value to the user's understanding of the meaning of the word. Moreover, replacing or changing existing examples in the preparation of DSLL2 seems unnecessary, considering that the authors are presenting the dictionary as a resource that reflects 150 years of the Slovenian language.[16] This is confirmed by Krek (ibid.: 147), concluding that this approach erased a great deal of evidence on the usage of words before 1991.

The *Dictionary of New Words of the Slovenian Language* (2012; DNWSL) was published even before DSLL2, and its authors to some extent used state-of-the-art lexicographic methods and included (whole-sentence) corpus examples, in addition to excerpts. As stated in the Introduction (DNWSL: 9), the main resource in the compilation of the dictionary was Nova beseda, a 318-million-word corpus of Slovene: [17]

> Based on authentic usage, attested in the 300-million-word Nova beseda corpus, 5,384 dictionary entries consist of 6,512 senses and subsenses of newer words and multi-word units, coming from different domains.

12 http://www.dwds.de/

13 Other dictionaries, e.g. DCD, offer access to a corpus on their website, however they do not offer a simultaneous search in all the resources and aggregated display of hits.

14 DWDS does offer the option of limiting the hits to only selected sources.

15 As written in the Introduction to DSLL (1991: XXII), "[w]henever the texts didn't contain enough information, the excerpts were either taken from other resources or invented".

16 Marko Snoj 2nd November 2013 for STA: http://www.rtvslo.si/kultura/knjige/akademska-vojna-okrog-novega-slovarja/321592.

17 http://bos.zrc-sazu.si/s_beseda3.html

A close examination of the examples in DNWSL reveals that the absence of (good) examples in Nova beseda sometimes forced the lexicographers to obtain them from other corpora, especially from 1.2-billion-word Gigafida corpus. Although this may not be problematic, it does bring into question the above cited methodology of headword list compilation, especially at entries such as *bandži skok* ('bungee jump'):

> **bándži skòk --** skôka in skóka m (ô, ȍ ó; o̩)
> skok v globino, pri katerem je skakalec pripet z dolgo elastično vrvjo; skok z elastiko: Obnaša se kot frkolin, ki se pred tovarišijo postavi z bandži skokom, ko se privezan na elastično vrv vrže z mostu v globel **E ↑bungee (jumping) in (↑)skòk**

The example provided above is a (slightly) modified sentence from the Gigafida corpus. It is noteworthy that the Nova beseda corpus does not contain a single hit for *bandži skok* (even Gigafida has only five). The example is thus attesting the use of a word for which we do not even know how it got into the dictionary. In addition, the dictionary's focus on newer words, which tend to have lower frequency in corpora, means that examples are used merely for attestation purposes, as they do not bring any added value to the understanding of the meaning.

A more systematic and corpus-driven approach has been used in the compilation of the Slovene Lexical Database (Gantar et al. 2012; SLD). The SLD contains 2,500 entries with 152,996 examples, so on average over 61 examples per entry. All the examples are whole sentences and were taken from the Gigafida corpus (Logar Berginc et al. 2012). The examples in the SLD have not been modified in any way, as the selection of examples for a lexical database differs from the selection of examples for the dictionary. Namely, the examples in the SLD also have the potential to become good dictionary examples, with only a few modifications needed. The SLD is particularly important for Slovenian and international lexicography because of the methodology used in its compilation. Namely, several methods combining lexicographic work with automatic extraction of data (including examples) have been developed and tested, and represent a basis for the compilation of the *Dictionary of Modern Slovene Language* (DMSL) and its database (see Section 5).

4 CHARACTERISTICS OF A GOOD DICTIONARY EXAMPLE

The most frequently mentioned characteristics of good dictionary examples are naturalness or authenticity, typicality, informativeness, and intelligibility. Naturalness means that the example appears natural, i.e. like the one you would expect

to encounter in actual language use. It is for this reason that the naturalness of dictionary examples is often associated with authenticity, which is ensured by obtaining examples from corpora, collections of authentic texts, something that has become a standard practice in modern lexicography. It should be pointed out that dictionaries compiled before the corpus lexicography era already contained examples from authentic texts (e.g. the *Oxford English Dictionary*), or at least excerpts based on authentic texts (e.g. DSLL). However, many of those dictionaries adopted a practice of formulating or inventing examples based on lexicographers' intuition. Overreliance on one's intuition has been brought into question by the findings of corpus studies (e.g. Sinclair 1991; Hunston and Laviosa 2001), which is particularly relevant when selecting examples for general monolingual dictionaries.[18]

Similar to the principle of naturalness is the principle of typicality – examples must show typical usage of the word in terms of context, syntax, phraseology and the like. State-of-the-art corpus tools can already significantly help lexicographers with this task, as they can be used to identify common, and typical, grammatical relations, collocations, and even colligations of the word (e.g. predominant number of the word in a particular collocation).

An informative example brings added value to the entry, predominantly in terms of offering additional help to the user in understanding the definition. In addition, examples attest the information in the definition, and contextualise the use of the word in a particular sense or subsense. The informativeness of an example is also affected by the number of examples in the entry. Electronic media offer the possibility of including a high quantity of examples, although lexicographers should always be concerned with whether each additional example offers anything new to the entry. On the other hand, as Frankenberg-Garcia (2012; 2014) pointed out, several corpus examples can sometimes be even more useful than the definition.

Intelligibility of a dictionary example is achieved by selecting examples that do not contain complex syntax or rare or specialised vocabulary. Examples should also not be too long. All this will help users focus on the word and the relevant surrounding information, and reduce the amount of mental effort needed to process it all. Still, certain features are often difficult to avoid; for example, rare and "more demanding" words are often used together with other rare words, which means the lexicographer needs to select such examples to fulfil the criteria of naturalness and typicality. While examples should not be too long, they must also not be too short, especially if a dictionary is to be used for encoding purposes where the users require as much contextual information as possible.

18 This is less true of dictionaries for non-native speakers, as, according to Atkins and Rundell (2008: 456), many pre-corpus English dictionaries for non-native speakers contained many good dictionary examples, which looked authentic but were not.

Form has become an important characteristic of a dictionary example; whole sentences are found in more and more dictionaries, even in general monolingual dictionaries for native speakers, which until a few decades ago used only excerpts or (very) short examples. There are two main reasons for this development: firstly, studies have shown that excerpts and similar short examples, taken out of sentences, seem abstract and unnatural (see e.g. Williams 1996). Secondly, in printed dictionaries, shorter examples are preferred due to spatial constraints, and the rise of digital media, especially the online medium, has done away with this limitation.

A separate and very important topic in example selection is ideological perspective, as examples reflect the ideology of the dictionary, i.e. reality as seen by lexicographers. Lexicographers use examples to convey information that could not be included in the definition because it is either too complex or ideologically too explicit (cf. Meschonnic 1991; Béjoint 2000; Epple 2000; Schutz 2002; Gorjanc 2004; 2005; 2012). Consequently, examples are an element of dictionary microstructure which offers the clearest reflection of social values, and relatedly, the values of the dictionary team (Gorjanc 2014). Analysing vocabulary related to homosexuals in DSLL, Gorjanc (2014) shows how social stereotypes can be presented in a dictionary as acceptable or part of the norm. Problems with ideological changes can also be observed in examples in DSLL2 (Krek 2014a: 145-147). It is therefore vital that lexicographers selecting examples are aware of their non-neutral role, and thus sensitive to social values and socially responsible (Béjoint 2000: 124).

Finding an example that meets all the above mentioned criteria is far from easy. Although nowadays lexicographers have very large corpora and consequently many potential dictionary examples at their disposal, it is often the case that they find sentences that meet two criteria, even three, but very rarely those that meet all the criteria of a good dictionary example. In fact, candidates for dictionary examples could be grouped on a scale from bad, more bad than good, reasonably or potentially good, and good; good candidate examples are those that can be used in a dictionary without any modifications. But, as mentioned above, such examples are less common, and there are more potentially good examples, i.e. examples that need minor modifications. However, if the decision is made to include modified examples in the dictionary, what about the principle of authenticity? Will the dictionary, or dictionary examples, still be considered corpus-based? As argued by Atkins and Rundell (2008: 458), the choice between invented and authentic examples is often misleading, because it does not reflect actual lexicographic practice. Even corpus-driven dictionaries like COBUILD include modified examples, although it should be stressed that the COBUILD lexicographic team tried to avoid modifying the examples as much as possible (Fox 1987).

Most common forms of example modification are shortening or omission of irrelevant parts, such as relative clauses or interjected clauses, simplification of complex syntax, and replacement of rare words or phrases with more common ones, or marked vocabulary with less marked vocabulary. Shortening is probably the least contentious practice, and is generally needed to meet the criterion of informativeness, as sentences often contain parts that can be deemed redundant or irrelevant, if not provided with more context. This is the case with *na primer* ('for example') in the sentence for the headword *anonimnost* ('anonymity'):

> *Jane Austen, na primer, je živela v popolni **anonimnosti**.*
>
> Jane Austen, for example, lived in total **anonymity**.

On the other hand, the simplification of complex syntax and replacing certain words can significantly affect the naturalness or typicality of an example. There are cases when replacing words cannot be avoided, for example proper names need to be replaced with pronouns or generic names to avoid offending individuals (e.g. Janez Novak; 'John Smith') or words that may offend particular social groups need to be replaced with more neutral ones. However, even this is not always straightforward, especially if the person in question is a public person or the name is closely related to the context of the word that is exemplified. The corpus example for *mojstrsko* (masterfully) would not have had the same informative value, and would also not appear natural, if *Christiano Ronaldo* had been replaced with a generic name such as Janez Novak ('John Smith'):

> *Izid polčasa in tudi končni izid je z **mojstrsko** izvedenim prostim strelom postavil Cristiano Ronaldo.*
>
> The half-time score, and also the final result, was decided by Cristiano Ronaldo's **masterfully** taken free kick.

The frequency and extent of example modification also depend on the target users of the dictionary. Examples for a dictionary for non-native speakers, or a dictionary for younger native speakers who are still developing their language proficiency and possess a smaller vocabulary, will be subjected to modification much more often than examples intended for dictionaries targeted at adult native speakers.[19] The decision about modification of examples should also be driven by expected use. For example, if the dictionary is supposed to help the users with both decoding and encoding, then the examples must remain as natural and typical as possible.

A special form of example modification is language correction. If we find a good corpus sentence with a missing comma, can we insert a comma and include the

19 Even Atkins and Rundell (2008) limit their approval of example modification almost solely to dictionaries for non-native speakers.

example in a dictionary? And if a sentence contains a misspelled word, a word in the wrong case, or an incorrect word order? Correcting spelling and some other minor mistakes may seem trivial, but the line between a minor and a major mistake can be very subjective. Some lexicographers may consider the replacement of a longer phrase or wording as completely acceptable, even though this is very close to inventing the examples altogether. To sum up, it is good to adhere to the principle of giving priority to finding good corpus sentences that do not need any modification, and only when such sentences cannot be found do we look for sentences with (minor) language errors that can become good dictionary example if these are corrected.

5 METHODS OF IDENTIFYING GOOD DICTIONARY EXAMPLES

Identifying (good) dictionary examples is a very laborious and potentially time-consuming, and thus expensive, process. One reason is that finding an example that meets all the criteria is very difficult. Moreover, corpora are getting larger and larger, which in most cases means a bigger selection of example candidates for the lexicographer, but also a greater number of examples to analyse. Thirdly, examples are a key microstructural element, and can be found under many different parts of the entry, such as senses, subsenses, compounds, phrases, collocates etc. All this means that the lexicographer needs to search for a lot of examples in a large amount of data for each dictionary entry.

There are two methods of identifying good dictionary examples: manual and semi-automatic. When using the manual method, the lexicographer can use the sort, filter and other functions of the corpus tools. Additional help is provided by the division of examples according to collocations and grammar relations. In the Sketch Engine[20] corpus tool (Kilgarriff et al. 2004) this option is offered by the Word sketches feature.

In the semi-automatic method, a tool for identifying good dictionary examples, such as GDEX (Good Dictionary Examples; Kilgarriff et al. 2008), offers a selection of candidate sentences to the lexicographer who then selects the most appropriate ones. GDEX (see also Section 5.1) ranks corpus sentences according to characteristics such as length, whole-sentence form, sentence complexity, presence or absence of rare words, email addresses or web addresses, and so on. Many of these characteristics are indirectly related to characteristics of a good dictionary example, such as typicality, informativeness and understandability. The characteristics can be divided into mandatory and less/more desired. The former are those

20 http://www.sketchengine.co.uk/

that the example must include; if only one characteristic is scored as negative, the example is ranked to the bottom of the list of candidate sentences. For the less/more desired characteristics we set the points added or deducted for each characteristic, and the example ranking is determined by a total sum of points from all the characteristics.

The main difference between the two methods is how much time they take, as the semi-automatic method is much quicker than the manual one, without being any less reliable (Kosem et al. 2012b; 2013b). In modern corpus lexicography, the semi-automatic method is thus replacing the manual method, especially in projects that involve the compilation of dictionary databases which include more examples than the dictionaries based on them (e.g. CDD).

6 EXAMPLES IN THE DICTIONARY OF MODERN SLOVENE LANGUAGE

This section discusses the treatment of examples in the proposed DMSL, including their identification and way of recording them in the dictionary database, and the differences between the examples in the dictionary database and the dictionary. The section concludes with a discussion on the different options of example presentation offered by digital media.

6.1 Identifying and recording dictionary examples

Identification of examples with the GDEX tool is part of a semi-automatic method called the automatic extraction of lexical data (AELD; Kosem et al. 2012b). This includes the automatic extraction of data (grammatical relations, collocates and examples, as well as certain information on the headword and also suggestions for labels), via Word sketches in the Sketch Engine, using an API script, taken directly from the corpus and put into a dictionary-writing system (DWS). In DWS, the data is then examined, selected and edited by the lexicographers.[21] This still provides the lexicographers with enough information for a thorough analysis and entry compilation. Experience on the SLD project has shown that a lexicographer using this method inspects a similar amount of examples, often even more of them, than by using a combination of semi-automatic and manual methods with the corpus tool (i.e. analysing word sketches). One of the advantages of AELD is that it dispenses with a lot of tedious copying and pasting of

21 A similar method has already been envisaged by Rundell and Kilgarriff (2011).

data between the corpus tool and the DWS. Another advantage is the quicker, more dispersed and consequently more reliable analysis.

The key task of AELD in terms of dictionary examples is the preparation of GDEX configuration(s). The GDEX configuration developed in 2011 during the compilation of the SLD (Kosem et al. 2011) was quite successful in identifying good dictionary examples, as on average three out of 10 examples offered were considered good. However, the requirements of AELD are different; only the first X examples (usually three to five) are extracted and all are expected to be (potentially) good. In addition, the analysis of the initial configuration for Slovene has pointed to significant differences between the quality of examples across word classes. This is why the decision was made to devise a different configuration for each word class. The procedure was done in two steps: first, the initial configuration for each word class was devised by analysing (good) examples in the SLD. New versions were then developed by adjusting the values of different classifiers and evaluating the results by comparing them with those of the previous configuration. The procedure was repeated until the GDEX configurations that provided the most satisfactory results were obtained. Importantly, the procedure also enabled us to devise several new classifiers that were not part of the original GDEX configuration. The classifiers used in AELD are:

- Whole sentence. Whole-sentence examples are given priority.
- No tokens with a frequency of less than three. This classifier seeks to exclude[22] examples with rare words, rare misspellings or corpus noise.
- Sentence must be longer than seven tokens. We seek to avoid examples that are too short, as they are often lacking context. The principle is that it is easier to shorten longer examples than search for new ones.
- Sentence must be shorter than 60 tokens. Only very long sentences are excluded, longer sentences can always be shortened.
- Lemma must not occur more than once. This important classifier excludes examples with a repeated headword, as such examples are normally less understandable and informative.
- Sentence must not contain web or email addresses.
- Optimal length (between X and Y tokens). While the classifiers for minimal and maximum length exclude sentences that are either too short or too long, this classifier awards points to sentences with the length in a given range. The most frequently used range is 15-40 tokens, but it depends on word class. The analysis of good examples in

22 The word exclude is used here because the algorithm ranks such examples so low that the lexicographers in most cases do not see them.

LBS, which was part of the preparation of the first version of GDEX for Slovene (Kosem 2012), has shown that the average length of examples for adjective entries is 28.64 tokens, for nouns 27.03 tokens and for adverbs 27.39 tokens.

- Rare lemmas. The classifier penalises sentences for each rare lemma. The frequency limit determining what is rare is determined considering the size of the corpus.
- Tokens longer than 12 characters. The classifier penalises sentences for each token that meets this criterion. This is because analysis has shown that tokens longer than 12 characters are in most cases non-words or corpus noise.
- Number of punctuation marks (commas excluded). The classifier penalises sentences that contain more punctuation marks than a set value. Commas are not included in the count as they are addressed by a separate classifier.
- Number of commas in the sentence. The classifier penalises sentences with more than three commas, as analysis has shown that such sentences are often more complex and thus less likely candidates for good examples.
- Tokens beginning with a capital letter. The classifier penalises sentences containing tokens with capital letters, and the main purpose is to complement the classifier penalising proper names.
- Tokens with mixed symbols (e.g. letters and numbers). Another classifier that helps identify, and penalise, non-words and corpus noise in the sentences.
- Proper names. The classifier penalises sentences containing tokens that are tagged as proper names. The penalty is awarded for each proper name in the sentence.
- Pronouns. The classifier penalises every pronoun in the sentence. The classifier is particularly important for sentences with several pronouns, as these often require a lot of additional context and are thus less understandable.
- Position of lemma in the sentence. The classifier penalises sentences in which the headword occurs outside of a given range in the sentence. For example, for the verb headwords it was determined that much better example candidates are sentences in which the headword does not occur at the beginning of the sentence (in the first 40% of tokens in the sentence).

- Stop list of sentence initial words. The evaluation of various configurations revealed that certain words at the beginning of a sentence are a good indicator of a bad candidate sentence. During the evaluation, a list of such words was devised. The list includes words such as *sledi* ('following'), *tovrsten* ('such'), *oboji* ('both') and so on, indicating that the sentence requires additional (preceding) context. The classifier penalises sentences beginning with one of the words on the list.
- Stop list of sentence initial multi-word units. The classifier is similar to the one above, penalising sentences beginning with a multi-word unit found on the previously devised list.
- Second collocate. One of the most important classifiers, awarding points to sentences containing the most typical collocates of a given collocation, indirectly detects colligational typicality. For example, the sentences containing the collocation *klavrn* + *podoba* ('miserable state') are awarded points if they also contain the verb *kazati* ('show'), which is a statistically important collocate of this collocation. Further analysis has also shown that such sentences also contain a longer syntactic pattern *kazati klavrno podobo česa* ('indicate miserable state of sth')
- Levenshtein distance. An algorithm[23] that measures similarity between strings, in our case sentences. If the classifier finds two similar or even the same sentences, it sends one of them (the one with the lower score) to the bottom of the list of candidate sentences.

Most of the differences between configurations for different word classes can be observed between the settings of individual classifiers, although differences in classifiers can also be observed (e.g. an additional classifier for the position of the lemma in the sentence is found only in the configurations for verbs). Each sentence receives a score between 0 and 1, indicating a total of all classifier values (as mentioned above, classifiers are attributed weights according to their significance in comparison with other classifiers). The GDEX tool then ranks the candidate sentences from the highest to the lowest score, and this determines which top X examples are exported with the AELD method.

Each example should include metadata about the text, such as year, source, author, title and so on. This ensures example traceability and offers different possibilities of example visualisation in the dictionary. It is never good to consider only the needs of a particular dictionary, as searching the corpus for missing information is a long and time-consuming process (this is true not only for examples but also for the other parts of the dictionary). A good indication of the benefits of example metadata is seen in the updating of the dictionary:

23 http://en.wikipedia.org/wiki/Levenshtein_distance. This measure was recently replaced by the Jaccard similarity coefficient (https://en.wikipedia.org/wiki/Jaccard_index).

if one wants to replace older examples with newer ones, it is possible to use the information on the year in which the text was produced to identify all the examples that were produced before a certain time. Example metadata can also be useful in the detection of ideological examples in the entry. Thus, when extracting with GDEX, we should pay particular attention to cases when most of the examples in the entry or under a particular sense are from a single source or only a few sources (cf. the analysis of examples for *pederastija* in Gorjanc 2014).

6.2 Examples in a dictionary database vs examples in a dictionary

The discussion on identifying and recording examples also needs to consider the relationship between a dictionary and its database, and also the role of the dictionary archive (Figure 1). The procedures described in this paper are especially relevant for DMSL, but since the compilation of this dictionary involves undertaking Slovene language description from scratch, a large proportion of the data obtained with corpus analysis (including examples) could be used in the compilation of other dictionaries. A particular example could thus be used in different dictionaries; in dictionaries for adult native speakers it can be used without any modifications, while in dictionaries for younger speakers the example can be slightly modified, e.g. by shortening or replacing rare words with more frequent ones that these users are likely to be more familiar with. Due to such potential multi-purpose nature of dictionary examples, all extracted corpus sentences and their metadata need to be archived in their original form, as found in the corpus.

An archive of extracted corpus sentences also makes possible analysing the number and type of modifications made to these sentences when turning them into dictionary examples. The findings of such analyses can then be used to improve the configurations used in their extraction. Even bad or irrelevant sentences that are part of automatic data extraction and need to be excluded from the database should be archived, as analysing their characteristics can also help improve the configurations for extraction. A similar approach was already used when developing the first version of GDEX for Slovene (Kosem et al. 2011); the parameters of the classifiers in the test configuration were improved with an analysis of the examples that were selected (good) or not selected (bad) during the evaluation. In addition, the role of dictionary data in the development of language technologies for Slovene should not be forgotten. In short, the planning stage of a dictionary project should devote a considerable amount of time to considering the various types of data in the dictionary database and the ways they will be recorded. From this perspective, any dictionary, even a

general dictionary, is merely one of several products that can be derived based on the database.

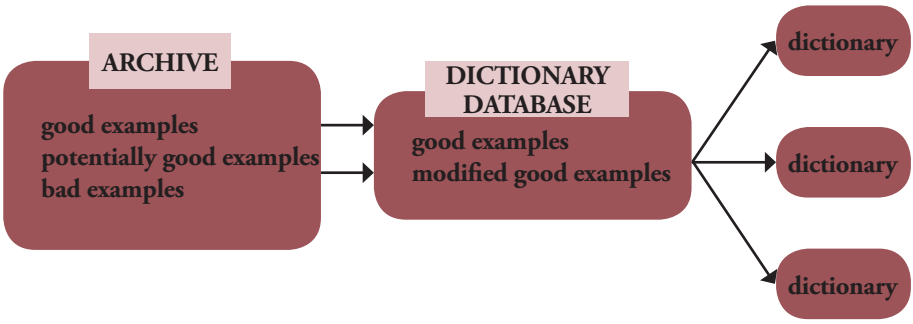


Figure 1: Examples in an archive, dictionary database and dictionaries.

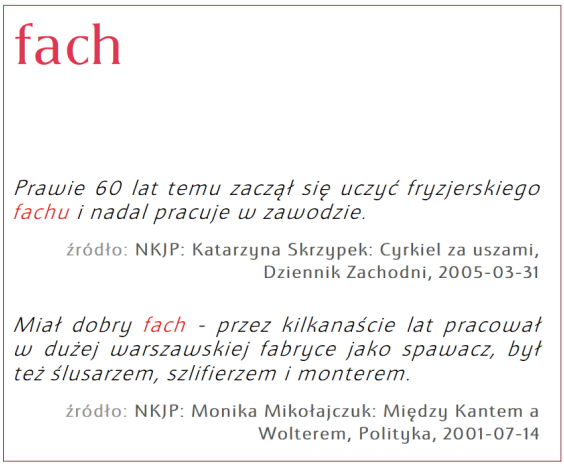
A dictionary database contains (much) more information than any dictionary based on it, which means that lexicographers can spent a great deal of their time recording information that might not end up in the dictionary. As such, planning of the dictionary database should follow two principles: a) automating as many (routine) lexicographic procedures as possible, and b) ensuring that every single lexicographic decision is recorded and utilized. Consequently, the use of methods such as AELD is more or less mandatory, as without them it is difficult to imagine the successful compilation of the database (and a dictionary based on it) in a time frame that would satisfy funders as well as dictionary users. Let us consider the benefits of using automation on a very basic task, namely typing a headword and its word class in the dictionary entry in the database. Assuming it takes us on average five seconds to type these two types of information, we spend on this task 500,000 seconds for 100,000 entries, or little less than 139 hours. AELD writes these two types of information automatically, saving us nearly one person/month on the project. Much the same is true of lexicographic decisions: using manual analysis or analysis in a corpus tool, even if using a tool like GDEX, lexicographers must still examine many corpus sentences and decide whether each is a good dictionary example or not. But since the lexicographers only copy good or potentially good examples in the dictionary database, only such decisions can be archived. The AELD method makes it possible to record or track every single decision: the identification of a good example (a corpus sentence remains unchanged in the dictionary database, so it is the same as the final dictionary example), a potentially good example (the corpus sentence has been modified slightly when turning it into a dictionary example), and bad examples (the corpus sentence has been deleted from the database).

To additionally assist lexicographers with the identification of good examples, other methods such as crowdsourcing can be utilized. However, as good dictionary examples have to meet a combination of different criteria, it is difficult to imagine how such a task could be trusted to non-lexicographers. The answer is that it can be, if we are aware of the characteristics and limitations of crowdsourcing (see Čibej et al. 2015; Fišer and Čibej 2015). First and foremost, the tasks should be simple, mainly in the form of multiple-choice questions with options 'Yes', 'No' and 'I don't know'. In addition, the tasks should not focus on determining something abstract (e.g. the characteristics of a good dictionary example) or level of degree; questions such as *Is this a good dictionary example?* and *How good is this example?* are thus not suitable. Tests with crowdsourcing on examples from the SLD have shown that examples are very useful in tasks aimed at identifying incorrect information (e.g. when the use of the headword and its collocate in the example does not match the identified grammatical relation) or at assigning collocates and their examples to different senses and subsenses.

6.3 Visualizing examples

Lexicographic work with examples does not, or should not, conclude with the recording of good examples in the database or/and the dictionary. This is because presentation is very important if examples are to achieve their purpose. Research studies in the visualisation aspects of (electronic) dictionaries, although still rare in lexicography, indicate that visualisation plays a key role in the readability and retention of the dictionary information (Nesi 2011). Considering that the examples occupy a fairly large, if not the largest, share of text in any dictionary, suitable visualisation and presentation of them is obviously vital.

One of the techniques used to assist users in reading dictionary examples is highlighting the headword. Especially in modern dictionaries that often contain (longer) whole-sentence examples, it is useful to direct users' attention to the headword, i.e. the part of the entry with information more relevant to their needs. In most cases such highlighting is found in the form of bold text, while in electronic dictionaries a different colour is also used (Figure 2). Italics are rarely used for highlighting, mainly because in most dictionaries examples are already offered in italics, and so this option seems less effective (see Figure 3). Highlighting is also used to point to typical collocations, compounds, multiword units and phrases (Figure 4). However, it is definitely recommended to test any visualisation and presentation solution on the target users, preferably before publishing the dictionary.



fach

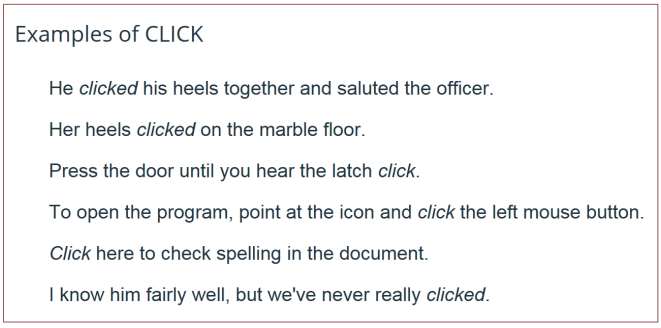
Prawie 60 lat temu zaczął się uczyć fryzjerskiego fachu i nadal pracuje w zawodzie.

źródło: NKJP: Katarzyna Skrzypek: Cyrkiel za uszami, Dziennik Zachodni, 2005-03-31

Miał dobry fach - przez kilkanaście lat pracował w dużej warszawskiej fabryce jako spawacz, był też ślusarzem, szlifierzem i monterem.

źródło: NKJP: Monika Mikołajczuk: Między Kantem a Wolterem, Polityka, 2001-07-14

Figure 2: Red headword highlighted in examples (CDP).



Examples of CLICK

He *clicked* his heels together and saluted the officer.

Her heels *clicked* on the marble floor.

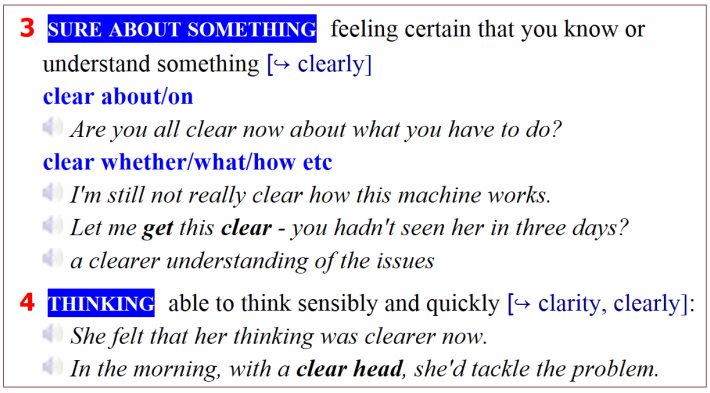
Press the door until you hear the latch *click*.

To open the program, point at the icon and *click* the left mouse button.

Click here to check spelling in the document.

I know him fairly well, but we've never really *clicked*.

Figure 3: Highlighting the headword in examples using italics (MWOD).



3 **SURE ABOUT SOMETHING** feeling certain that you know or understand something [↪ clearly]

clear about/on

Are you all clear now about what you have to do?

clear whether/what/how etc

I'm still not really clear how this machine works.

*Let me **get** this **clear** - you hadn't seen her in three days?*

a clearer understanding of the issues

4 **THINKING** able to think sensibly and quickly [↪ clarity, clearly]:

She felt that her thinking was clearer now.

*In the morning, with a **clear head**, she'd tackle the problem.*

Figure 4: Highlighted collocations and phrases in examples (*Longman Dictionary*) [24]

24 http://www.ldoceonline.com/

As already mentioned, it is useful to have as much metadata as possible on each example in the dictionary database. Although such metadata can be shown to the users, dictionaries rarely include it – one exception is the *Comprehensive Dictionary of Polish* (Figure 2) – for a simple reason: metadata such as source, author(s) or title of the text from which the example comes from are referential and suggest/require direct copying from the source, taking away from the lexicographers the option of making any modifications. Another reason against showing example metadata is its non-essential nature; it would take up precious space on the screen and can distract the users' attention from the main purpose of the examples, namely showing the use of the headword in a particular sense.

The principle of informativeness also limits the lexicographers in the number of examples they can provide under each element of the entry. Even with that in mind, one can quickly end up with several examples per sense, subsense, syntactic structure or collocation, which can cause problems with visualisation/presentation. A good solution is to show only a certain number of examples, offering additional examples on a click (Figures 5 and 6). More and more online dictionaries have also started to offer links to corpus hits, undoubtedly a very useful feature for (advanced) users.

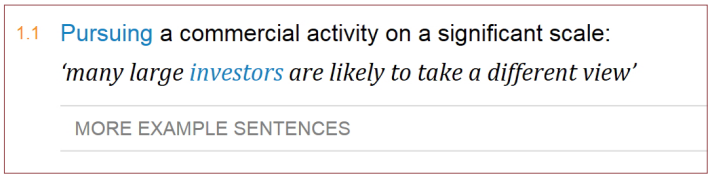


Figure 5: A link to show additional examples (more example sentences in Oxford Dictionaries).

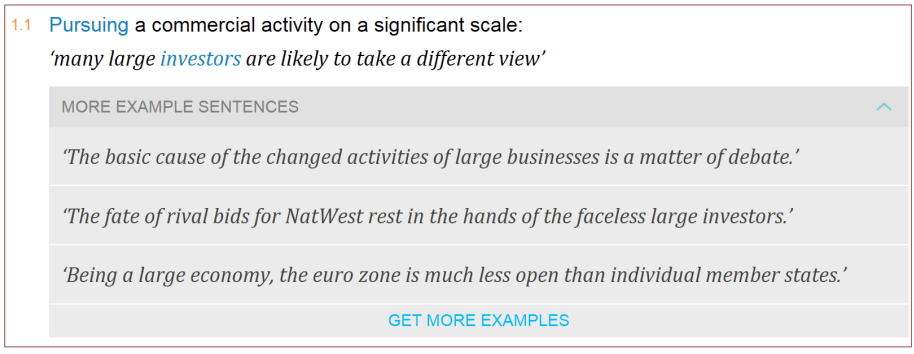


Figure 6: Additional examples revealed (Oxford Dictionaries).

7 CONCLUSION

Examples require a great deal attention when planning a dictionary. The instructions given to lexicographers thus need to clearly delineate the characteristics of good examples, including concrete cases of good and bad practice, and the role of ideology. It is also paramount to use or develop tools that facilitate consistency in adhering to these characteristics. In addition, examples of allowed modifications should be prepared, as well as a suitable system for archiving sentences in the form they are extracted from the corpora. DMSL will be an important resource for the development of language technologies for Slovenian, which means that the database should include as many examples (and their metadata) as possible.

The aim to include as many examples as possible necessitates the use of semi-automatic methods of example extraction from the corpus. Not using such methods can prolong the compilation of the dictionary to such extent that the examples need to be replaced before the work is even completed. This is the rationale behind using the AELD method that we propose for identifying and recording examples in DMSL, and which represents a new approach to lexicographic analysis. Based on the experience gained during the SLD project, a similar method has already been used in the compilation of a collocations dictionary for non-native speakers of Estonian (Kallas et al. 2015).

An important task for the lexicographic community is to keep conducting studies on how, when and in what ways dictionary users consult examples and what kind of examples are most useful to them. The findings of such studies will enable further improvements to the procedures used for example selection, and the techniques used to present them in dictionaries.