Make My (Czechoslovak Word of the) Day

Michal Škrabal[1], Vladimír Benko[2]

[1] Charles University, Institute of the Czech National Corpus,
Panská 7, 110 00 Praha 1, Czech Republic
[2] Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics,
Panská 26, 811 01 Bratislava, Slovakia
E-mail: michal.skrabal@ff.cuni.cz, vladimir.benko@juls.savba.sk

Abstract

Our paper introduces an experiment aimed at creating a database to be used as the source for a *Word of the Day* (*WotD*) application. Using a database of translation equivalents derived from a Czech-Slovak parallel corpus as a point of departure, semi-automated procedures are described that would preprocess the raw data so that the size of the lexicon to be processed manually is minimized. A by-product of this experiment is a list containing Czech to Slovak translation equivalents of differing levels of similarity, which could be an interesting source of information for Czech and Slovak contrastive studies.
In the last chapter the lexicographical application of acquired data is described. The criteria for selecting individual headwords remain an open question at the moment. Personally, we lean towards a combination of different aspects so that the final selection is as diverse and user-attractive as possible. The intended microstructure of the *WotD* dictionary entry is also presented. Its first peculiarity is the dual metalanguage, making it two explanatory dictionaries in one rather than a translation dictionary. Secondly, the content of the entries is closely related to the digital-born and corpus-based nature of the dictionary. Thus, some elements presented in traditional explanatory dictionaries are reduced or completely omitted in our microstructure – while others are highlighted.

Keywords: Word of the Day; translation equivalent; Czech; Slovak; *Treq* database

1. Introduction

Many online dictionaries and other lexicographic/didactic resources have their *Word of the Day* (*WotD*), a feature that on a daily basis focuses on a chosen lexeme, giving users a wide range of varied information about it. For example, Merriam-Webster's *WotD*[1] presents the profile of the selected word every day. It makes reference to the pronunciation of the expression, its definition, a brief commentary on the origin of the word and connection with other, related words and, eventually, two or three examples of its usage, most often from current media, sometimes also from older literary works. To make *WotD* even more interactive and entertaining, it also contains numerous links to additional materials concentrated on the Merriam-Webster web portal (such as *Test Your Vocabulary*, *Word Games*, *Trending Now*, *Words at Play*, etc.).

[1] https://www.merriam-webster.com/word-of-the-day.

Another example, *One Hungarian Word a Day* (*OHWaD*)[2], aims at a different target group: being written in English, it is primarily intended for L2 students of Hungarian. At the beginning they are asked to guess the meaning of the selected word from three possibilities, whereupon the correct English equivalent is revealed. Subsequently two or three example sentences are given as well as a short glossary of semantically close words and phrases with their English counterparts. This way students learn six new words from Monday to Saturday, whereas Sunday is dedicated to revision in the form of a quiz: students are supposed to choose the correct equivalent for six newly learned words and to use each word in the made-up Hungarian sentence.

Another concept hidden under a similar name can be found in, for example, the Polish project *Słowa dnia*[3]. These "words of the day" are based on the relative frequency of words in daily newspapers that is clearly higher than their frequency in the comparative period of the previous year (cf. also Meriam-Webster's *Word of the Year*[4] based on the frequency with which each word has been searched for in the dictionary in the past year). Of course, frequency may be one of the criteria for selecting such "prominent" words (see also chapter 4 below), nonetheless, our project is closer to the first two projects mentioned above.

Since, at least to our knowledge, there is no such project for either Czech or Slovak, we thus propose a simple database to help generate individual parts of such a series for either of these languages. It would be a rudimentary automated system open to extra modules that could facilitate lexicographers' work and utilize the corpus data (that are available for both languages in abundant volume) as much as possible. Besides, it combines a modern, quantitative approach with traditional lexicographical practice (definitions taken from older printed dictionaries, etymological information, etc.) and incorporates the long-standing and very popular tradition of so-called "linguistic columns" (called *jazykové koutky* in Czech / *jazykové kútiky* in Slovak) into a lexicographical project.

2. The data

Though a bilingual Czech to Slovak dictionary (Horák et al., 1979; cf. also Gašparíková & Kamiš, 1967; Nečas & Kopecký, 1964) was available in machine-readable form, we decided not to use it for this project, mainly for two reasons. Firstly, its paper version was published four decades ago and therefore does not reflect recent developments in either the Czech or Slovak lexis, especially after the political changes in our societies since 1989. Secondly, as it had been compiled in the pre-corpus era, many translation equivalents are not sufficiently attested, or are even simply wrong (cf. Ripka &

[2] https://www.catchbudapest.com/one-hungarian-word-day.

[3] http://slowadnia.clarin-pl.eu.

[4] https://www.merriam-webster.com/words-at-play/word-of-the-year-2018-justice.

Skladaná, 1980). Moreover, we *could* use a resource that is much more up-to-date, with translation equivalents attested in a parallel corpus and supplemented with frequency data.

2.1 The *Treq* database

The *Treq*[5] application serves for querying the Czech to foreign language(s) dictionaries that have been automatically created based on data derived from the *InterCorp* parallel corpus (Čermák & Rosen, 2012). This parallel corpus also includes a Czech-Slovak component (Nábělková & Vavřín, 2018) that currently (in version 11) comprises the following text types:

- fiction (the so-called *Core* [of the corpus])[6] – 10.5 million tokens;
- legal texts of the European Union from the *Acquis Communautaire* corpus – 23.3 million tokens;
- proceedings of the European Parliament dated 2007-2011 from the *Europarl* corpus – 14.8 million tokens;
- movie subtitles from the *Open Subtitles* database – 7 million tokens.

The overall size of the whole *InterCorp* v11 corpus is more than 1.7 billion running words / 2.14 billion tokens[7], of which more than 45.4 million running words / 56.2 million tokens accounts for a Czech-Slovak component (i.e. less than 3%). Nevertheless, this amount of data is sufficient for our purposes.

Access to the extracted data[8] is mediated by the *Treq* online search interface (http://treq.korpus.cz/). The application provides a list of all translation candidates of a given word (or even multi-word expression) found in *InterCorp* that are, by default, sorted by decreasing frequency. The more often the equivalent of the search term occurred compared to other equivalents, the higher the probability that it is plausible.

2.2 The *Treq* dump format

Besides the online access, the *Treq* database was available for use in the framework of our experiment in a simple three-column text format, containing the *frequency*, *Czech*

[5] The acronym *Treq* stands for *Translation Equivalents.*

[6] Only fiction texts have been manually corrected in terms of OCR and sentence alignment. All other texts were processed automatically only. For the list of tools used, see http://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze9#acknowledgements.

[7] For information about the exact composition of the corpus and the size of its components, see http://wiki.korpus.cz/doku.php/en:cnk:intercorp. For general information about the InterCorp project, see Čermák & Rosen (2012) or Rosen (2016).

[8] For detailed information about the automatic processing of data, see Škrabal & Vavřín (2016).

word, and *Slovak word*, respectively. Lists for both lemmatized and raw word forms derived from the four basic *InterCorp* components were provided in eight separate files, with total word counts as shown in Table 1.[9]

Treq component	**Word forms**	**Lemmas**
Core	433,962	198,346
Acquis	808,812	438,023
Europarl	716,703	348,963
Subtitles	489,324	292,231

Table 1: *Treq* source data.

All files were sorted by descending frequency. The first 20 lines of the two *Acquis* files are shown in Table 2.

Lemmas			**Word forms**		
Freq	**cs**	**sk**	**Freq**	**cs**	**sk**
807,038	.	.	806,355	.	.
764,487	,	,	752,456	,	,
473,280	a	a	478,142	A	a
345,762	v	v	343,364	)	)
343,458	)	)	254,050	V	v
249,537	(	(	249,578	(	(
215,721	na	na	207,633	na	na
190,750	být	byť	141,178	se	sa
190,027	článek	článok	124,192	O	o
144,811	se	sa	112,538	1	1
140,941	s	s	112,514	nebo	alebo
132,084	nařízení	nariadenie	92,848	;	;
130,075	který	ktorý	90,316	Článek	Článok
125,418	o	o	88,897	"	"
116,925	z	z	87,046	2	2
113,391	nebo	alebo	85,123	:	:
113,238	komise	komisia	84,227	S	s
112,617	1	1	83,258	Č	č
112,416	stát	štát	76,867	pro	pre
109,491	společenství	spoločenstvo	70,417	-	-

Table 2: *Treq* source data (*Acquis*).

[9] We are grateful to Martin Vavřín for kindly providing us with this data.

The source of the data is easily recognizable by the nouns present in the list: *článek/článok* 'article', *nařízení/nariadenie* 'regulation' or *komise/komisia* 'commission', clearly indicating the EU legal discourse.

We decided to use the lemma files for further processing only. The data in the following text are based on the *Acquis* file.

3. Preprocessing

Czech and Slovak are languages belonging to the West Slavic group that are very close and to a large extent mutually intelligible. There exist, nonetheless, some differences at the phonetic, orthographic and lexical levels[10] that are targeted by our *WotD* project.

It is obvious that the list of candidate entries should not contain only identical or "similar" lexical items. They should predominantly consist of equivalents that are "sufficiently different". As the resulting list will have to be eventually validated by a linguist, the preprocessing should aim to eliminate as many "similar" words as possible, so that the list to be processed manually is not too long. The frequency information is naturally another indication to take into account.

3.1 The pipeline

The preprocessing was performed by means of simple Linux tools: *egrep* utility for regex-based filtrations, *sed* batch editor for character substitutions, and *cut* and *paste* utilities for column manipulations.

The processing pipeline consisted of the following steps:

- adding rank numbers to lemmas;
- removing items containing non-alphabetical characters (66,568 lines removed);
- removing items containing uppercase letters (mostly proper names and abbreviations; 35,736 lines removed);
- removing single-letter items;
- removing items with identical source and translation (42,660 lines removed) – here is the respective regex trick:

  ```
  egrep -v "[[:space:]]([[:alpha:]]+)[[:space:]]\1$" input >output
  ```

- deleting diacritics that denote the lengths of vowels (*á* > *a*, *é* > *e*, etc.), as well

[10] See e.g. Sokolová, Musilová & Slančová (2005: 5), who refer to F. Uher's and M. Sokolová's older research from the 1980's. According to them, there is a formal and semantic agreement between the Czech and Slovak texts in 38% of lexemes, a partial agreement even in 46%, while 16% are problematic in terms of communication. Out of the 500 most frequent lexemes in Czech and Slovak, 230 (46%) were completely identical, 154 (31%) were partially identical and 116 (23%) were completely different.

as the palatalization of consonants (e.g., *ď* > *d*, *ľ* > *l*, etc.); removing identical items after this filtration using the same regex trick (4,899 lines removed);

- deleting all vowels; removing identical items after this filtration (19,397 lines removed).

At this point, we still had 95,572 candidate translations that could finally be reduced by applying a frequency threshold. After some experimentation, we decided to set it to 100. The sizes of all four resulting lists are shown in Table 3.

Treq component	**Lemmas (original list)**	**Lemmas (filtered list)**	**%**
Core	433,962	1,867	0.43
Acquis	808,812	5,007	0.62
Europarl	716,703	3,517	0.49
Subtitles	489,324	910	0.19

Table 3: Preprocessed data

The next table shows the first 20 (out of more than five thousand) *WotD* candidates filtered from the *Acquis* list.

Rank	**Freq**	**cs**	**sk**	**Rank**	**Freq**	**cs**	**sk**
12	132084	nařízení	nariadenie	45	44348	moci	môcť
16	113391	nebo	alebo	49	37488	vzhledem	keďže
19	112416	stát	štát	51	34376	ohled	zreteľ
27	84698	evropský	európsky	59	31823	smlouva	zmluva
29	75711	tento	toto	62	29039	země	krajina
30	72069	český	č	65	27721	jenž	ktorý
32	62315	tento	táto	66	27326	odstavec	odsek
38	53562	pro	na	70	25467	všechen	všetok
39	52703	být	sa	71	24945	zejména	najmä
43	44726	být	by	73	23678	jiný	iný

Table 4: *WotD* candidates based on the *Acquis* list.

The Rank column contains rank values from the original list, which makes it apparent how many words have been deleted during the step-by-step filtration (i.e., lemmas with rank 1-11, 13-15, 17-18, 20-26, etc., were deleted). The resulting list still contains a certain amount of noise (e.g., the item ranked 30 is most likely a result of different lemmatization policies for abbreviations being used by the various taggers), yet even among the first 20 items, we can find very good *WotD* candidates. In general, lists preprocessed in the described way not only can save a lot of time for linguists, but can put the whole enterprise into the "doable" category.

3.2 The data filtered out

As an interesting by-product of the above procedure, we also got three lists of translation equivalents that are equal or "reasonably similar". These data can be of some interest not only to linguists in the areas of contrastive studies, language typology, phonology, etc., but also to translators – it is a known fact that translation between close languages is straightforward only in a deceptive sense. Some examples are given in Tables 5 and 6; however, this is beyond the purview of our current paper.

Rank	Freq	cs	sk
8	190750	být	byť
42	46147	příloha	príloha
69	25842	případ	prípad
75	23441	příslušný	príslušný
82	22021	hospodářský	hospodársky
89	19945	den	deň
100	18508	oblast	oblasť
102	18290	třetí	tretí
115	16531	při	pri
147	13068	měnit	meniť
151	12697	činnost	činnosť
166	11496	předpis	predpis
180	10390	změna	zmena
189	10040	před	pred
196	9767	část	časť
200	9308	agentura	agentúra
207	9042	veřejný	verejný
220	8607	stanovit	stanoviť
247	7757	další	ďalší
260	7411	přístup	prístup

Table 5: Most frequent translation equivalents differing in quantity of vowels and soft consonants only

Rank	Freq	cs	sk
9	190027	článek	článok
10	144811	se	sa
13	130075	který	ktorý
17	113238	komise	komisia
20	109491	společenství	spoločenstvo
28	79802	pro	pre
33	59929	směrnice	smernica
40	52366	opatření	opatrenie
41	49881	rozhodnutí	rozhodnutie
50	37213	mít	mať
55	33263	soulad	súlad
68	25978	podle	podľa
72	23803	informace	informácia
76	23010	podmínka	podmienka
80	22439	muset	musieť
86	20407	výrobek	výrobok
97	18814	svůj	svoj
99	18720	žádost	žiadosť
103	18080	společnost	spoločnosť
104	18001	společný	spoločný

Table 6: Most frequent translation equivalents differing in combination of vowels and soft consonants only

4. Lexicographic application

The *WotD* application is meant to be the first step in a broader *WotD* project, ideally one involving both Czech and Slovak lexicographers – as both Czechs and Slovaks form the target group of users. Confronting the dual view of the same topic would certainly be beneficial to both nations, which once lived together within one country. The Czech and Slovak languages would once again stand side by side, as they did before. While they are mutually intelligible to the older generation who remembers the Czechoslovak federation, for the youngest generation this is far from being the case – quite often using English as a mediating language.

4.1 List of headwords

The question of choice of words for the *WotD* project is crucial and deserves an elaborated conception. Nonetheless, whatever criteria are chosen, the point is that preselection of the candidates is taken care of by a computer, and a lexicographer only revises automatically generated drafts of entries. Our application generates a further editable draft version of the given entry, relying primarily on corpus data (frequency, most common collocations, exemplification using the *GDEX* tool (Kilgarriff et al. 2008), etc.), complemented by a lexicographic description taken from existing dictionaries and by other features. Such a draft would be subsequently edited by a lexicographer who would also write a brief commentary – a usage note or even an essay (the Czech lexicographer would comment on a Czech word whereas the Slovak lexicographer would comment on a Slovak word – or, occasionally, even vice versa). As these feuilletons on the various linguistic subjects are rather popular in both countries, we believe a broad audience would become interested in the project. After all, the public can be actively involved in it – e.g. by commenting on individual words on the project website, by voting for the most popular word(s) or for words to be processed in the future, or in other ways.

The criteria for selecting individual headwords remain an open question at the moment. The pipeline described above eliminated formally similar words from the candidate list. However, even these may appear in the final inventory – although being words common to both languages, they are still potentially different in their use (including cases of false friends), frequency, etc. However, the largest group of words will naturally be those specific for one of the languages – with the most common equivalent(s) in the second language, including pairs that are the source of the linguistic humour[11]. Personally, we lean towards a combination of different aspects so that the final selection

[11] In the Czech environment it has long been believed that Czech *veverka* 'squirrel' is called *drevokocúr*, literally 'tree-tomcat', in Slovak. Such a word, however, does not exist in Slovak at all, as a formally similar word *veverička* is used. See Nábělková (2008: 219-232) for the description of this inter-language myth in detail.

is as diverse (both semantically and grammatically) and user-attractive as possible. The aim is to educate the audience in an engaging form: we want readers to realise on the one hand the interconnection of these two languages (lexicon inherited from a common Slavic basis, mutual reciprocal loanwords, commonly used internationalisms), on the other hand their diversity, deepening after the break-up of Czechoslovakia in 1993.

4.2 The microstructure of the WotD entry

With regard to the microstructure of individual *WotD* entries, the whole project has at least two specifics. The first one is the dual metalanguage – Czech and Slovak, making it, de facto, two explanatory dictionaries in one rather than a translation dictionary. Mutual equivalents here serve only as a secondary means to emphasise a contrastive nature. A top-down layout with a vertical partition seems to be ideal: the left half will be reserved for the Czech part of the entry, the right half for the Slovak part, while the individual elements of the microstructure will be horizontally aligned side by side.

In addition, it is a born-digital project that would result in an electronic dictionary that can be augmented in the event of public interest by any number of items. We take processing a set of 365 dictionary entries as a suitable beginning, provided that a new entry is published daily for the time span of one year. The inventory would then gradually expand and in the final stage it would cover, albeit in an unbalanced way, the whole alphabet. In fact, there would be two lists of entries – a Czech one and a Slovak one, both easily searchable. A close connection between the dictionary and corpora in the form of numerous links is commonplace.

The content of the entries will also be closely related to the born-digital and corpus-based nature of the dictionary. Some elements presented in traditional explanatory dictionaries would thus be reduced or completely omitted in our *WotD* microstructure – while others would be highlighted. For example, in traditional dictionaries the lemma is most often followed by morphological/grammatical information. In *WotD*, the emphasis would be laid on frequency data and usage specifics (typical genre/text-type, communication situation, etc.). This should be demonstrated by some suitable examples which, in the spirit of the famous Firthian dictum “You shall know the word by the company it keeps” (Firth, 1957: 11), would illustrate the meaning(s) of the word, but also its creative alterations in specific texts (e.g. fiction) or in spoken language. The difference between the spelling and the pronunciation of Czech/Slovak words is not as large as in English, therefore the sound recording of the word could move from the heading of the entry to the exemplification part – and indicate, among other things, different semantics and usage within spoken and written language (which is, at least in Czech, close to diglossia – Bermel, 2014). In addition, the exemplification section should include a direct link to the corpora concerned, providing additional examples to a potentially interested person.

The example part would be followed and supplemented by the lexicographer's commentary, the imaginary central part of the whole entry. It should be written in a popular, entertaining style and should aptly reflect the place of the given word in the lexical system of language, along with the differences from the second language. These may appear on the diachronic level, as a variance in the development of semantics and/or the use of the same word in both languages. Therefore, basic etymological information should be provided as well.

Only at the end of the entry can explanatory definitions from existing Czech and Slovak dictionaries be cited. Although this is the central part of the entry in traditional dictionaries, we perceive them instead as an interesting appendix providing a contrast to the modern, corpus-based approach to lexicography.

The microstructure of the dictionary entry is far from definitive; on the contrary, it is a mere suggestion that should underline the specificity of our project and which will need to be properly tested by compiling several sample entries.

5. Acknowledgements

This work has been, in part, funded by the Slovak VEGA Grant Agency, Project No. 2/0017/17. It was also supported by the European Regional Development Fund-Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734). During its creation we used the tools developed within the Czech National Corpus project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

6. References

Bermel N. (2014). Czech Diglossia: Dismantling or Dissolution?. In J. Árokay, J. Gvozdanović & D. Miyajima (eds.) *Divided Languages? Diglossia, Translation and the Rise of Modernity in Japan, China, and the Slavic World.* Cham: Springer.

Čermák, F. & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus, *International Journal of Corpus Linguistics* 13 (3), pp. 411-427.

Firth, J. R. (1957). *A synopsis of linguistic theory 1930-1955.* In J. R. Firth (ed.) *Studies in Linguistic Analysis*, Special volume, Philological Society. Oxford: Blackwell, pp. 1-32.

Gašparíková, Ž. & Kamiš, A. (1967). *Slovensko-český slovník.* Praha: Státní pedagogické nakladatelství.

Horák, G. et al. (1979). *Česko-slovenský slovník.* Bratislava.

Kilgarriff, A.; Husák, M.; McAdam, K.; Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the 13th EURALEX International Congress.* Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu

Fabra, pp. 425-432.

Nábělková, M. (2008). *Slovenčina a čeština v kontakte – Pokračovanie príbehu.* Bratislava: Veda.

Nábělková, M. & Vavřín, M. (2018). *Korpus InterCorp – slovenština, verze 11 z 19. 10. 2018.* Ústav Českého národního korpusu FF UK, Praha 2018. Accessed at: http://www.korpus.cz [17 July 2019].

Nečas, J. & Kopecký, M. (1964). *Slovensko-český, česko-slovenský slovník rozdílných výrazů.* Praha: Státní pedagogické nakladatelství.

Ripka, I. & Skladaná, J. (1980). Česko-slovenský slovník. *Slovenská reč* 45(6), pp. 364-372.

Rosen, A. (2016). InterCorp – a look behind the façade of a parallel corpus. In E. Gruszczyńska & A. Leńko-Szymańska (eds.) *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora.* Warszawa: Instytut Lingwistyki Stosowanej, pp. 21-40.

Sokolová, M., Musilová, K. & Slančová, D. (2005). *Slovenčina a čeština (Synchrónne porovnanie s cvičeniami.* Bratislava: Filozofická fakulta Univerzity Komenského v Bratislave.

Škrabal, M. & Vavřín, M. (2017). The Translation Equivalents Database (Treq) as a Lexicographer's Aid. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference.* Leiden: Lexical Computing.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

http://creativecommons.org/licenses/by-sa/4.0/

.