An Author's Dictionary: The Case of Karel Čapek

František Čermák
Charles University Prague

After a brief reminder of the long tradition of manually-based author's dictionaries, the possibility of a dictionary based on a full corpus and verified in a number of aspects against a large corpus has re-emerged. Specifically, the plan of Karel Čapek's dictionary and its realisation is being discussed and its final shape shown, having a number of new, hitherto unused features. The Dictionary, being in fact split into four separate ones, is accompanied by the full Čapek's corpus on a CD where a lot of additional information can be found.

1. Goals and history

The idea that the entire vocabulary of an important person, a man of letters preferably, should be captured and described in its entirety and serve then as a model and tertium comparationis for others is old and there are many early specimens of this to be found. These include classical authors of Antiquity such as Homer, or internationally famous figures such as Shakespeare, Dante or Goethe, or, to meet a more national demand, local authors important for native speakers of a particular language only, such as Petöfi, Ibsen or Mickiewicz, though, in some cases, only specific (usually famous) books have been described in this way. However, most of these have been compiled manually, not even (sometimes) offering an exhuastive coverage, presenting their data as word lists with various statistics or/and, sometimes, in the form of a concordance. As such, these books do not make a very good reading, offering tedious long lists and a number of figures. Only the very last of these dictionaries, from the second half of the 20th century, could have been compiled, though in a very limited manner, using first computers (see Arnold 2006, Mattausch 1990, Roelcke 1994).

2. A modern corpus-based author dictionary: The case of Karel Čapek

It seems that a really modern author's dictionary based on a corpus and compiled using today's computer facilities has not been compiled, so far. Next to a large Goethe dictionary which is under making and which does not seem to be finished yet (*Goethe-Wörterbuch*, 1978 ff.), the only exception seems to be the dictionary of a Czech 20ieth century classic Karel Čapek (1880-1938), which has been published in 2007 (*Slovník Karla Čapka*, Dictionary of Karel Čapek, for short *Dictionary* in what follows). The choice fell on him not only because of his considerable reputation (having been widely translated, introducing, among other things, the word *robot* into the international vocabulary) but, primarily, because of his language, which is still, after some 60-80 years, surprisingly modern and excercising a huge influence on modern Czech language.

3. Logistics and preparation

As Karel Čapek's entire works have not been represented in the Czech National Corpus (*http://ucnk.ff.cuni.cz*), his published books had to be scanned and compiled into a special corpus. Since it has been the intention of the authors to publish, next to the dictionary itself and a number of additional materials, also his corpus with a corpus manager and browser on a CD, a lot of administrative paperwork including negotiating permission of his inheritors to do so had to be gone through, too. No dictionary can capture a mans language entirely, not being able to capture all of its syntagmatic features mostly. Acknowledging this, a deliberate move and offer has been made towards users by offering them the full corpus of Čapek to search, research and explore, should they feel so or should they need necessary contexts for some words or collocations, etc. Due to this research option and perhaps a kind of broadly speaking educational facility offered, the edition of the Dictionary is a bit more than a mere dictionary.

To be true to the idea that it should be only his language that was going to be represented and researched, some of Čapek's works have not been included, though they are published on the accompanying CD separately (all of this to be published also as part of the Czech National Corpus eventually, see above). These included specifically those texts that were written together with his brother or interviews with president T. G. Masaryk where it was impossible to distinguish between texts of these two distinguished men in the published texts. As an extremely prolific author, Čapek has written several distinctive types of text, namely fiction, poetry, journalism by which he made his living on a day-to-day basis, though he also wrote on art and philosophy as well as on his gardening hobby, etc. For this reason, the bulk of his work has been divided into five major broad genres. It was the comprehensive edition of Čapek that has come out in the 1980ies that has been used as basis for the Čapek corpus. Alas, due to several spelling reforms and problematic zeal of subsequent editors, the language of the texts is not exactly the one the author used. Although our intention has been to have thus as formally unified texts as possible, this has not, in fact, been possible and many forms and words of Čapek's are still represented in more than one form.

Having obtained the author's texts in an electronic and unified form, a lot of work has gone then into tokenisation, tagging and subsequent lemmatisation. Obviously, a lot of specific technical work had to be done, too, both during the preparation of the corpus and the dictionary (see more M. Křen). It has to be pointed out that Čapek's language, despite the feeling of modernity it gives, is not exactly contemporary, however. This meant that no programmes, created recently for modern Czech corpora, could have been used directly and without problems. Hence, a series of quasi-manual corrections and unification, in which fourteen people in their various capacities and skills have been employed, the team being led by František Čermák.

The texts include fiction (37% of all texts, including drama and poetry), journalism (47,5%), technical texts (philosophy, aesthetics, culture and art, 5%), personal correspondence (10%) and translations (mostly of French avantgarde poetry), the bulk of his language belonging to the two first of these categories.

During his 22 years in which he wrote, i.e. since his first up to his last book, Karel Čapek has created a considerable bulk of texts. Those included in the Dictionary are made up of some 68,900 single-word lemmas or lexemes (including proper names) that are based on 2,676,688 text tokens (equal to 196,936 types); these have been used in cca 207,613 sentences, the average sentence thus consisting in slightly over 11 word forms (11,15). Should one include those texts that have been left out (see above), his vocabulary would obviously be somewhat larger. To give a comparison, the Goethe vocabulary is estimated now to be formed by some 90,000 lexemes, but Goethe wrote for more than twice the time (over 50 years) and his vocabulary, due to a different typological character, is rich in very many compounds, corresponding to collocations in Czech that are not counted as lemmas and, accordingly, as a dictionary's item. It is evident that any straightforward comparison is impossible.

4. The dictionary of Karel Čapek

Due to different types of information as well as of data, the final product has been split to appear in four separate dictionaries, namely as *Slovník* (Dictionary), *Slovník hapaxů* (Dictionary of Hapaxes), *Slovník proprií* (Dictionary of Proper Names) and *Slovník zkratek* (Dictionary of Abbreviations). Although the main, i.e. the first, dictionary is the most important, a brief explanation and description of all of these will be given in the following.

Slovník (i.e. the main alphabetical Dictionary itself) is a compromise between what the authors intended to publish and what was feasible, also because the book is meant to be both a specialized handbook for researchers and a book for a general reader. The solution, representing a compromise between many alternatives, is then based on the decision to include all lemmas having frequency 2 or higher which are accompanied by a number of features.

Each *lemma* is given a total frequency figure followed by six other figures indicating distribution of this total frequency in the *six genres* that have been recognized as significant (see below). These (being then sum of the total frequency) are represented by (1) fiction (in Czech *próza*), (2) drama, (3) journalistic texts (*publicistika*), (4) poetry, (5) professional literature (*odborná*), and (6) personal correspondence. This allows the user to arrive at his/her judgement on what the Present time and

that of Karel Čapek have linguistically in common and on what has changed since, as well as on the author's preferences for use of specific lexemes in some genres only. In particular, this is to be seen in his use of international words. In fact, this can easily be done by the user himself, comparing data for Čapek with the reference representative contemporary corpus SYN2005 which is available on the web (2005 designating the year of its release, i.e. some 80 years later or more). It has to be noted that both corpora, SYN2005 and that of Karel Čapek, have been tagged and lemmatized in the same way, hence they are perfectly comparable. Next to that, the Dictionary is provided with five additional, optional and, hopefully, useful features (A-E, see also examples below). These include

a brief *annotation of meaning* (definition) of lesser known lemmas in italics. These could represent technical, international, time-bound and now obsolete words, a number of foreign words but also his own lexical creations, for which he is famous. There are no less than at least twelve languages which he knew or quoted from, this being a considerable source of problems during automatic tagging and lemmatisation, too;

annotation of non-Czech lemmas, whether those academic (such as Latin *augur*) or just borrowed and quoted (the English *attention*, German *ausgerechnet*) by an abbreviation of the language of origin (in parentheses);

words and lemmas that were never used by the auhor in isolation (or very seldom) are given a special "plus" sign following them (such as *absurdum+ (lat.)* to be found in the Latin collocation *ad absurdum* only).These, obviously, are used as parts of larger phrasemes, multiword terms, etc.;

if a lemma is preceded by an asterisk (*), it is a signal for the user that it occurs in the author's texts in a *significantly higher* degree than is its today's frequency in the reference 100-million corpus mentioned above. In this way, the whole vocabulary of Karel Čapek has been statistically checked against this corpus and differences higher than 0.001 have been noted. In practice, this means that there is a 99.9% certainty that the difference in occurrence found for such words as **čelo* (forehead, forefront), or **demokracie* "democracy" is not due to chance and a student of the author is here given a systematic and solid indication of Čapek's lexical preferences;

many lemmas are provided with a set of specific and typical *collocations* found in the Čapek's corpus (introduced by a bullet sign ●). A set of bigrams serving as a basis (i.e. after figures, punctuation marks, proper names and abbreviations have been deleted) have gone through calculations to determine the thresholds for exclusion of many peripheral items. These consisted in MI-score figures (the threshold for exclusion chosen being lesser than 4), phi-score and log-likelihood (the threshold being less than 10) while, finally, all collocations with the overall frequency smaller than 3 have been excluded, too. No need to say that a considerable amount of experimenting with various scores has been undertaken and the final decisions are based empirically on this with the aim to exclude uninteresting collocations on the one hand (mostly those of grammar words) and accidental collocations on the other hand. Thus, a set of some 5000 collocations has been arrived at that had, at the same time, a mininal summary rank of all the three association measures. As an additional benefit for the user, the asterisk is used also here to mark those collocations in this list that are, on the basis of χ^2 measure, statistically more significant than those in the reference 100-million corpus (SYN2005), the level of probability being 0.001. It is often collocations that signal readily important phenomena of the outside world. Thus, one of the early collocations of, for example, *koncentrační tábor* (concentration camp) has been recorded also by Karel Čapek. This dictionary has 34,740 lemmas.

The second of the four dictionaries is *Slovník hapaxů* (Dictionary of Hapaxes) presenting alphabetically all lemmas that Karel Čapek has used only once, which, in his case then, are his specifix hapax legomena. The phenomenon, well-known from the study of old authors since Antiquity and being a source of vexation for lexicographers since they were unable to say a thing about these text loners, is, in fact, not much different here. As some of these are not to be found elsewhere and their derivative nature and, hence, relation to other words, is problematic and often uncertain, they may represent a challenge even decades after the author's death. There are 19,585 of these recorded.

The fact that Karel Čapek was a broadly and widely educated man, a journalist, a public figure and a man of culture having rich contacts with a number of leading personalities of the contemporary Europe and that he has traveled extensively in many countries is reflected in the unusually high

number of proper names, many from classics but also from among his contemporaries. *Slovník proprií* (The dictionary of Proper Names) offers, thus, next to names of people, those of places that have been of interest for him, and that have been often, through his newspaper articles, related to contemporary topics. Thus, one finds here such lemmas as *Cervantes, Cézanne, Cicero, Cordóba*, but also *Hitler* or *Lenin* though these have only a marginal frequency, the highest being *T.G. Masaryk*, the first president of Czechoslovakia and his personal friend (286 occurrences). Also here lemmas are quantified so that their distribution in the six genres mentioned above is given. There are 12,442 lemmas here.

The last of the dictionaries included is a small Slovník zkratek (Dictionary of Abbreviations) being more or less a formal appendix, having 314 items and being presented along the lines observed in the preceding dictionary. Yet, on second thoughts, even here one can find an important record of the pre-war reality and its important entities and institutions.

A Sample of Slovník

	Total frequency	Fiction	Drama	Journalism	Poetry	Proffesional	Correspondence
absurdum+ *(lat.)*	4	0	0	3	0	1	0
abundance	3	0	0	3	0	0	0
hojnost, *nadbytek*							
abúzus	2	0	0	2	0	0	0
nemírné užívání, *nadužívání*, *zneužívání*							
***aby**	10077	2817	269	5014	13	294	1670
ac *(velš.)*	4	3	0	1	0	0	0
acaulis *(lat.)*	2	0	0	1	0	0	1
acta+ *(lat.)*	4	1	0	3	0	0	0
actes *(fr.)*	2	0	0	2	0	0	0
acti *(lat.)*	2	0	0	2	0	0	0
ve spoj. *laudator temporis acti velebitel minulých dob*							
actu+ *(lat.)*	3	1	0	0	0	2	0
ve spoj. *in actu v průběhu*							
***ač**	106	12	2	61	2	7	22
…							
***čelo**	383	206	25	127	5	4	16
● *čelem vzad mnout si čelo přemnout si čelo rozpálené čelo svraštělé čelo svraštit čelo vraštit čelo zpocené čelo*							
čemeřice	10	0	0	10	0	0	0
čenich	7	4	0	3	0	0	0
čenichat	5	2	0	3	0	0	0
čep	6	2	0	3	0	1	0
čepec	13	10	0	3	0	0	0
● *pod čepec*							
čepeček	10	9	1	0	0	0	0
čepel	12	8	0	4	0	0	0
● *ohnivá čepel*							
***čepice**	97	62	1	28	2	0	4
● *placatá čepice úřední čepice vysoká čepice*							
čepička	14	8	0	5	0	0	1
● *kožená čepička*							
čepobití	3	3	0	0	0	0	0
zvukový signál ohlašující počátek nočního klidu, večerka							

The dictionary is an attempt to capture, in a dictionary form, the author's vocabulary in its totality. However, the compilers have undertaken an extensive research into his vocabulary, too, and the book offers some other useful additional parts, some of them being more technical in nature (linguistic), some aimed at a wider readership. Thus, there is a study of a foremost literary expert on Karel Čapek (Jiří Opelík, *Poznámky nelingvistovy*) to be found here, followed by an extensive analysis of his language (by a group of authors) where all major aspects are scrutinized and amply illustrated (*Slovník Karla Čapka: jeho lexémy a nominace*), especially many of his numerous innovations consisting not only in new words but also in new meanings and collocations. The book is provided with a number of appendices including a statistical survey of many features of his language against the background of other authors and, also, a valuable list of insightful and often witty remarks, aphorisms and quotations (*Myšlenky, aforismy a výroky Karla Čapka*), some of which have long been in use before this book, presenting the author as an important, influential and highly original thinker commenting on his time. An extensive bibliography is added, too.

5. Some open problems and concluding remarks

The Dictionary of Karel Čapek is also an attempt to document a period of the first half of the 20th century through the language and views of one of its prominent figures and writers; its particular value may be appreciated in comparison with other similar works. Within the framework of both modern lexicography and corpus linguistics, its particular usefulness is to be seen in its being anchored, especially for the sake of reference, in two corpora, that of the author himself and a large contemporary one. Thus, one may study the language of the period through one of the contemporaries in an exhaustive and objective way for the first time (which cannot be said of numerous, selective and therefore subjective previous studies written in a traditional way, preferred by literary people), one may also view it in its movement and change, though, last but not least, one may also study the time from a much broader cultural and political perspective.

References

[Anonymous]. *Concordance to Shakespeare*, 1787. London.

Arnold, B. (2006). "Night Watches on the Computer: Creating an Author's Dictionary with Computational Means". Literary and Linguistic Computing 21. 5-14. *http://llc.oxfordjournals.org/cgi/reprint/21/suppl_1/5.pdf.*

Barlett, J. (1984). *Complete Concordance or Verbal Index to the Words, Phrases and Passages in the Dramatic Works of Shakespeare*. New York: St. Martins.

Bessinger, J. (1969). *Concordance to Beowulf*. Ithaca: Cornell University Press.

Blanc, L. G. (1852). *Vocabolario Dantesco ou Dicctionnaire critique et raisonné de la Divine Comèdie de Dante Alighieri*. Leipsic: J. A. Barth.

Burger, A. (1957). *Lexique de la langue de Villon*. Genève: Droz.

Cunliffe, R. J. (1910). *A New Shakespearean Dictionary*. London: Blackie and Son Limited.

Čermák, F. (2001). *Jazyk a jazykověda. Přehled a slovníky*. Praha: Karolinum.

Čermák, F., M. Křen et al. (2004). *Frekvenční slovník čestiny.* Praha: NLN.

Čermák, F. et al. (2007). *Slovník Karla Čapka*. Praha: NLN.

Davies, N.; Gray, D. et al. (1979). *Chauser Glossary*. Oxford: Clarendon Press.

Desfeuilles, A. et P. (1900). *Lexique de la langue de Molière*. Paris: Hachette.

Dyce, A. (1902). Glossary to the Works of W. Shakespeare. London.

Galton, H. (1989). "The theory of verbal aspect and tense illustrated for czech by Karel Čapek's *Bajky a podpovídky*". *International Journal of Slavic Linguistics and Poetics* 35/36. 51-64.

Goethe-Wörterbuch. Stuttgart: Akademie der Wissenschaften der DDR, Akademie der Wissenschaften in Göttingen und Heidelberger Akademie der Wissenschaften.

Grzybek, P.; Stadlober, E. (2003). "Zur Prosa Karel Čapeks. Einige quantitative Bemerkungen". In Kempgen, S.; Schweier, U.; Berger, T. (eds.). *Festschrift für Werner Lehfeldt zum 60. Geburtstag*. München: Sagner. 474-488.

Hartmann, R. R. K.; James, G. (1998). *Dictionary of Lexicography*. London: Routledge.

Heinemann, M. (1990). "Schprachkomik bei Karel Čapek. Probleme ihrer Wiedergabe im Deutschen". *Zeitschrift für Slawistik* 35. 72-75.

Karel Čapek a český jazyk, 1990. Uspoř. F. Štícha, PedF UK, Praha.

Křen, M. (in print). *Compilation of the Dictionary of Karel Čapek*.

Kusse, H. (2005). "Im Krieg mit den Molchen: Stereotype der Aggression bei Karel Čapek". In Berwanger, K.; Kosta, P. (eds.). *Stereotyp und Geschichtsmythos in Kunst und Sprache*. Frankfurt am Main: Peter Lang. 631-646.

Lexicon Homericum (H. Ebeling, ed.). Lipsiae, 1885.

Lexicon Novi Testamenti Graeco-Latino-Belgicum. Pasor, 1690.

Lexicon Sophocleum (F. Ellendt, ed.). Berlin, 1872. (Přetisk 1965, Hildesheim.)

Livet, Ch.-L. (1895-97). *Lexique de la langue de Molière, comparée à celle des écrivains de son temps*. T. 1-3. Paris.

Lockwood, L. (1907). *Lexicon to the English Poetical Works of J. Milton*. New York.

Makin, M.; Toman, J. (eds.). *On Karel Čapek. A Michigan Slavic Colloquium*. Michigan: Ann Arbor.

Matějka, L. (1992). "The registers of Čapek's Czech". In Makin, M.; Toman, J. (eds.). *On Karel Čapek. A Michigan Slavic Colloquium*. Michigan: Ann Arbor. 51-57.

Mattausch, J. (1990). "Das Autoren-Bedeutungswörterbuch". In Hausmann, F. J.; Reichmann, O.; Wiegand, H. E.; Zgusta, L. (eds.). *Wörterbücher. Dictionaries. Dictionnaires, Ein internationales Handbuch zur Lexikographie.* Berlin: De Gruyter. 2, 1549-1562.

Mattausch J. (1990). "Textlexikographische Aspekte im Autorenwörterbuch (am Beispiel des Goethe-Wörterbuchs)". In Goebel, U.; Reichmann, O.; Barta, P. I. (eds.). *Historical Lexicography of the German Language*. Lewiston: Edwin Mellen. 2, 713-733.

Olminskij, M. S. (1937). *Ščedrinskij slovar*. Moskva.

Petõfi Sándor életművének szókészlete. Szerkesztette J. Soltész Katalin et al. Budapest 1973-1987. (4 díly)

Ramsay, R. (1936). *Mark Twain's Lexicon*. // Missoury Studies in English XIII.

Roelcke, T. (1994). "Autorenlexikographie". *Lexicographica* 10. 1-20.

Scabert, I. (1972). *Shakespeare. Handbook*. Stuttgart.

Schmidt, A. (1874-5). *Shakespeare-Lexicon: A Complete Dictionary of All the English Words, Phrases, and Constructions in the Works of the Poet*. 4th ed. Berlin.

Short, D. (1990). "Linguistic authenticity in Karel Čapek's Conversations with TGM". In *T. G. Masaryk (1850-1937), vol. 3 (Statesman and cultural force)*. London. 178-199.

Slovar jazyka Puškina, 1956-1961. Ed. V. V. Vinogradova, Moskva.

Słownik języka Adama Mickiewicza, 1962-1983. Ed. K. Górski, S. Hrabec, Wrocław.

Spevack, M. (1973). *The Harvard Concordance to Shakespeare*. Berkeley.

Těšitelová, M. (1948). "Frekvence slov a tvarù ve spise *Život a dílo skladatele Foltýna* od Karla Čapka". *Naše řeč* 32. 126-130.

Těšitelová, M. (1990). "Karel Čapek a jazyk". *Slovo a slovesnost* 51. 192-200.

Těšitelová, M. (1955). "Poznámky ke slovní zásobě v románu Karla Čapka *Život a dílo skladatele Foltýna*". *Naše řeč* 38. 297-307.

Turunen, A. (1979). *Kalevalan sanat ja niiden taustat*. Lappeenranta.

Umbach, H. (1986). "Individualsprache und Gemeinsprache. Bemerkungen zum Goethe-Wörterbuch". *Zeitschrift für germanistische Linguistik* 14. 161-174.

Wiegand, H. E. (1986). "Bedeutungswörterbücher oder sogenannte Indices in der Autorenlexikographie? Die Eröffnung einer Kontroverse". In Wiess, W.; Wiegand, H. E.; Reis, M. (eds.). *Textlinguistik contra Stilistik? - Wortschatz und Wörterbuch - Grammatische oder pragmatische Organisatione von Rede?* Tübingen: Niemeyer. 163-169.

Wimmer, G.; Altmann, G. (1999). "Review article: On vocabulary richness". *Journal of Quantitative Linguistics* 6 (1). 1-9.

Zielinski, M. (1981). "Probleme der Übersetzung Čapeks ins Deutsch. Die Rolle der Verbalaspekte in Čapeks Werk *Zahradníkův rok* und ihre Widerspiegelung in der Übertragung von Grete Ebner-Eschenhaym". *Zeitschrift für Slawistik* 26. 859-863.