Author Dictionaries Revisited: Dictionary of Bohumil Hrabal

František Čermák and Václav Cvrček
Institute of the Czech National Corpus, Charles University Prague

With a view to continue the line of author dictionaries, started by that devoted to Karel Čapek (2007), a second dictionary, basically following the first, has been compiled, namely that of Bohumil Hrabal (2009), an influential and major figure of the contemporary literary scene. The idea to have more of comparable and corpus-based dictionaries of this type that would ultimately enable comparison and through the prism of some of the best masters of the language to view the Czech language in development, has been made possible only recently, with the existence of corpora and thanks to techniques developed by corpus linguistics. A number of new lexicographic and computational features, never used before (with the exception of K. Čapek's dictionary), have been tried verifying options how to best put into practice general theoretical ideas, such as when finding best collocations that could be included in the dictionary.

1. Goal and Framework

To publish a second author dictionary, based on a corpus, within a couple of years may seem an uninteresting repetition (see Čermák 2008), not being very original. Yet, following the model approach used for mapping the vocabulary of a prominent Czech writer of the first half of 20th century (Karel Čapek 1890–1938) and applying it to another, no less prominent and widely translated author from the second half of the century (Bohumil Hrabal 1914–1997) should be taken both as an attempt to find out the extent to which it can be explored more widely, and making it possible to compare both.

In a broader context, detailed knowledge about dictionary of any author represents a much needed information about idiolect, i.e. vocabulary of an individual which, obviously, does not have to be a famous writer. This idea brings us back to a somewhat forgotten scale lying behind possible types of dictionaries, starting with those one is familiar with and can buy at booksellers. Yet the scale is, at least in theory, sociologically much broader, including descriptions of a (A1) state of language related to a broad community, or (A2) its section (geographical, social or other), (A3) a profession or science, (B) an individual, whether a writer or somebody else, and, finally at least, (C) a single book, such as the *Bible*. Considered from a somewhat finer point of view, some of these types may be further 'diluted' or filtered to be used for specific purposes, such as text-books, etc. In an ideal state of a description of a language at a given time, there should be as many of these as possible, forming a basis for strategic decisions about the language in questions. Needless to say that such an ideal state has not been really achieved; there is, in fact, no consensus that this is really necessary, let alone in what proportions it could be done. Such an idea, futile for some, may get a further justification should one realize that doing this one does not describe language only, but also everything it represents, the cultural and social reality it reflects, while frequencies of lexemes do also signal frequency of outside phenomena and that may become concern of many other types of people. To just give a single example from the Dictionary of Karel Čapek (Čermák 2008), a prominent author between the wars, knowing the sheer number of occurrences where he mentions Lenin, Hitler and Stalin, let alone his views about these three monsters, says a lot both about him and his time.

2. Modern Corpus-Based Author Dictionaries: A Dictionary of Bohumil Hrabal

The old idea of mapping the entire vocabulary of an important person, a man of letters preferably, has been briefly recalled earlier (Čermák 2007) and, indeed, quite a number of such dictionaries for a number of languages have been compiled in past, though mostly

manually (for a brief survey of the most important ones from a different non-computer period, see Čermák 2007). However, the approach, corpora and computing possibilities have dramatically changed so that it is just difficult to consider these to be comparable to what is possible today. Corpus-based dictionaries offer not only much more in descriptive features that the authors settle on using but also possibility of using a virtually unlimited context when the corpus is consulted by the user, enabling, among other things, his/her own interpretation or, perhaps, correction of what is being offered and stated in the dictionary. Thus, an entirely new and much richer type of information may be offered in this new generation of author dictionary based on corpus.

Bohumil Hrabal stands comparison to no one in the Czech literature and his unique type of writing has won him admiration abroad, too, as his books have been widely translated, made into films and greatly appreciated by native Czech readers. Of course, it is not only this phenomenal man and his books as such that were the reason for the choice to make him object of another of this type of dictionary coverage. His language, which is unique in many respects (some of his books having been written in the form of a single sentence, etc.), does cover most of the 20th century, namely three of its major periods, i.e. both before the war and after it including the whole of the Communist period and some time after the downfall of Communism and return to a free society. Unlike other of his contemporaries he held many original views of his time and its features that are worth being recorded.

3. Preparing Data

Today, it is natural to publish such a dictionary in conjunction with the corpus it is based on a CD, since the book form is always limited in size at least while the corpus offers additional look-ups by an interested user or scholar. To be able to do this, full corpus of Bohumil Hrabal had to be assembled (thanks to Hrabal's last editor) and painstakingly lemmatized and tagged first. In this case, as the author has died only recently (1997) it turned out that some of a number of people, being co-inheritors of his, when asked to give their approval to use the author's books for the dictionary, adopted an evasive tactics, hoping to squeeze some money from this and not realizing that the dictionary was not, strictly speaking, a commercial project.

Having obtained copyrights for Hrabal's texts we had to scan and OCR his entire works to compile it into a corpus. Excluded were only those texts where the authorship was dubious (such as interviews with Hrabal which represent not only Hrabal's language).

When this has been sorted out and the word list being created, separate and parallel activities have been started including computing representation of lemmas in one or more of the four genres (shorter fiction, longer fiction, journalism and poetry), filtering out hapaxes, proper names, abbreviations, numbers etc., computing collocations against the background of a large contemporary standard corpus of Czech, writing definitions to some of the lemmas, etc. Next to this, data for an extensive study of the author's vocabulary has been started, too, and computations regarding some advanced statistical ratios for comparison with other corpora (most of these being found in appendices, i.e. next to the dictionary proper). This may sound simple and straightforward but that was not the case, things being complicated, just like in all similar and non-standard projects, requiring a lot of post-editing, revision, finding suitable policy how to tackle parts of the author's texts that he often reused elsewhere or the changing spelling, etc.

First drafts of Hrabal's dictionary showed that we face another specific problem of the author, namely duplicate parts of texts in his collected works. This was due to three factors, mainly. Firstly it is Hrabal's unique writing technique which became famous for his repeated re-use of some parts of his older texts (it is also referred to as a collage or montage of dialogues or monologues he witnessed etc.), but also many fictional characters re-appear in several works uttering the same or similar remarks (such as uncle Pepin). Secondly, Hrabal often re-wrote some of his texts being dissatisfied with their form or style (therefore there are e.g. three variants of Příliš hlučná samota (Too Loud a Solitude) in his collected works with one being in standard 'literary' Czech, another written using colloquial language and the third having changed into a piece of poetry). Finally, there are duplicates caused by official censorship of the communist regime which made him modify some texts. These texts were after the fall of the Communist regime republished in their original form for the first time or some parts of these texts were used in newer texts. In order to obtain precise figures (since duplicates can influence frequencies of some rare words or measures used for collocations) we had to deal with these duplicates by generating a list of lemmas and word forms that were influenced by this phenomenon. The decision has been made that all texts larger than 13 words appearing in his corpus at least twice were signed as duplicates if they did not share more than 3 different tokens (these differences being usually a matter of punctuation).

Going by the dates of publication, Bohumil Hrabal's writing activities spanned over 37 years (1937-1995) and were projected into a text made up of 2 051 398 tokens (including punctuation, i.e. if one counts all real texts together as one) and these, in turn, are represented by 47 482 lexemes.

The data thus prepared have then been made into the dictionary while in their raw form they have been, in the form of a tagged and lemmatized corpus, placed on CD to accompany the the book. The CD that is distributed together with the dictionary is supplied with a corpus browser Bonito as well making the editing a self-contained package that allows for more specific search and study that could not enter the book.

4. The Dictionary (and its Linguistic Research)

The book is structured into three parts: collective linguistic study of Hrabal's language and, specifically, of his vocabulary, several dictionaries and appendices. The *study* based on comparison of Hrabal' corpus with the reference 100 million word corpus of contemporary Czech SYN2005 consists of a description of phonological, dialectological, morphological and lexicological features specific to Hrabal's works. In some cases, the comparison with the subcorpus of Czech fiction in the reference corpus was needed in order to differentiate between features specific for the author and specific for the style or text type. This theoretical description of Hrabal's lexicon is divided into four parts: direct nominations (use of lexemes), pragmatic nominations, indirect nominations, i.e. phraseology and idiomatics, and collocations.

The *dictionary* has been split in five separate dictionaries, namely *Slovník* (Dictionary), *Slovník hapaxů* (Dictionary of Hapaxes), *Slovník proprií* (Dictionary of Proper Names), *Slovník zkratek* (Dictionary of Abbreviations) and *Frekvenční slovník* (Frequency Dictionary). 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) includes all lemmas with frequency 2 or higher (26,542 entries) which are accompanied by a number of features. Each lemma is given

a total frequency figure followed by four other figures indicating distribution of this total frequency in the main four genres (shorter fiction, longer fiction, journalism and poetry). Next to that, the Dictionary is provided with additional features. These include

(A) a brief annotation of meaning (definition) of lesser known lemmas. These are mainly technical, slang and dialectal words. Hrabal was very interested in philosophy and art theory to which he also often addresses his life-long commentaries.

(B) annotation of non-Czech lemmas, whether those academic (such as Latin or Greek) or just borrowed and quoted (mainly German, French and English) by an abbreviation of the language of origin (in parentheses);

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

(D) if a lemma or a collocation is accompanied by one or more asterisks (*), 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 of Czech SYN2005. In this way, the whole vocabulary of Bohumil Hrabal has been statistically checked against this corpus by chi-square measure and differences on the significance level of 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 **pivo* (beer), or **esesák* (colloq. member of the German *essesman*, i.e. Schutsstafel during the war) is not due to chance and a student of the author is here given a systematic and solid indication of Hrabal's lexical preferences;

(E) many lemmas are provided with a set of specific and typical collocations found in Hrabal'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, or, rather, transferred elsewhere) 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), log-likelihood (the threshold being less than 10) and phi-score while, finally, all collocations with the overall frequency smaller than 3 have been excluded, too. All the bigrams were sorted according to these measures and ranked. 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 minimal 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 chi-square measure, statistically more significant than those in the reference 100-million corpus (SYN2005), the level of probability being 0,001.

(F) the number of duplicate contexts in which the lemma occurs is signalised by the number in upper index. This value can be interpreted as follows: the word *jaro*^{6} has overall frequency of 173. But 6 of these occurrences are duplicates in a context larger than 13 words (see above).

The second of the four dictionaries is *Slovník hapaxů* (Dictionary of Hapaxes) presenting alphabetically all lemmas that Bohumil Hrabal has used only once, which, in his case then, are his specific hapax legomena. There are 12 754 of these recorded.

Third part, *Slovník proprií* (The Dictionary of Proper Names) offers names of people, places and things that were used in Hrabal's texts. There are 7642 lemmas here.

The fourth of the dictionaries included is a small *Slovník zkratek* (Dictionary of Abbreviations) being more or less a formal appendix, having 140 items and being presented along the lines observed in the preceding dictionary.
Both of these minor dictionaries capture some of the most typical figures and institutions of the period in which the author lived.

Last of the dictionaries, *Frekvenční slovník* offers 5000 lemmas of the author, ordered by frequency.

A Sample of *Slovník*:

Lemma	Total frequency	Shorter fiction	Longer fiction	Journalism	Poetry
*almara[3]	105	71	24	4	6
	● *otevřít*[1] *almaru*[3]				
*almárka[1]	38	29	6	1	2
	● *otevřít*[1] *almárku*[1]				
*alou	8	3	5	0	0
alpa	2	1	0	0	1
	lihový bylinný roztok určený k masážím těla				
alpakový+	2	1	0	1	0
	ve spoj. *alpaková lžička* lžička vyrobená ze slitiny mědi, niklu a zinku				
alpinka	2	1	1	0	0
	alpská, resp. vysokohorská rostlina				
alpinum	2	1	1	0	0
	umělá zahradní skalka s vysokohorskými rostlinami				
alpský	6	1	2	0	3
*als (něm.)	12	2	5	0	5
	jak				
alt	7	1	4	0	2
*altán[1]	24	22	0	1	1
*altánek	23	19	4	0	0
*altare+ (lat.)	3	0	3	0	0
	ve spoj. *ad altare Dei* na oltář boží				
altový[1]	4	4	0	0	0
aluminiový[1]	3	2	0	0	1
aluminium	2	0	0	0	2
*am (něm.)	12	5	2	0	5
amalgám	3	2	1	0	0
amant (fr.)	2	1	0	1	0
	milenec				

The dictionary is accompanied by several *appendices*. These are of two types; firstly, there are additional studies of statistical aspects of Hrabal's language as well as the description of duplicates in Hrabal's works, secondly, the book provides the reader with complete lists of Hrabal's phrasemes, metaphors, proverbs and interesting ideas and views.

5. Open Problems and Concluding Remarks

Though the idea of having an author's dictionary is an old one, computer-based dictionaries are rather new and offer, in a sense, new possibilities enabling comparison of one single man vocabulary with that of another, though he/she might not belong to the category of famous people. However, compiling this kind of dictionary opens new possibilities as well as problems. Methodologically, it is obvious that having more dictionaries of the type from various time periods offers a chance to study idiolects in a principled and objective way and follow their developments through time. However, this idea has not been pursued so far, though it might become important. In fact, the study of idiolects may seem a valuable contribution to the general study of a language, started from the bottom, so to speak. More specifically, this approach involves decisions to be taken related to the type of collocations to be included, as these can be easily recalculated, given the corpus that has been made available to the user.

It is obvious that, for practical reasons, the features included and presented here had to be weighed against other that have ultimately been paid less or no attention. Should more dictionaries along similar lines follow, somewhat different approach might be chosen in the quest for an optimal one.

References

Čermák, F. (ed.). (2007). *Slovník Karla Čapka*. Praha: NLN. (together with: T. Bartoň, R. Blatná, V. Cvrček, M. Hnátková, J. Kocek, M. Kopřivová, M. Křen, K. Kučera,V. Schmiedtová, M. Stluka, M. Šulc, P. Vondřička, M. Waclawičová) (=A Dictionary of Karel Čapek).

Čermák, F.; Cvrček, V. (eds.). (2009). *Slovník Bohumila Hrabala*. Praha: NLN. (together with: T. Bartoň, M. Hnátková, J. Kocek, M. Kopřivová, M. Křen, K. Kučera, R. Novotná, V. Schmiedtová, M. Stluka, D. Šrajerová, M. Šulc, M. Waclawičová) (= A Dictionary of Bohumil Hrabal).

Čermák, F. (2008). *An Author's Dictionary: The Case of Karel Čapek*, In E. Bernal; J. DeCesaris (eds.). *Proceedings of the XIII Euralex International Congress*, Institut Universitari de Linguistica Aplicada Universitat Pompeu Fabra, Barcelona 2008, 323-332.