Towards Semi-Automatic Dictionary Making
Creating the Frequency Dictionary of Hungarian Verb Phrase Constructions

Júlia Pajzs and Bálint Sass
Department of Language Technology. Hungarian Academy of Sciences

The paper describes the lexicographical aspects of creating a frequency dictionary by a semi-automatic process. The bulk of the work is made by task specific software. The output of the program is then manually checked, corrected and filtered. The result is a collection of the most frequent Hungarian verb phrase constructions (VPCs), illustrated by corpus examples. This is a corpus driven dictionary, based on the 187,6 million word synchronic Hungarian National Corpus (http://corpus.nytud.hu/mnsz) which was analyzed by a series of programs. Its output is a set of XML format draft entries, which were then hand validated and edited by lexicographers. The dictionary contains the most frequent Hungarian verbs along with their most typical syntactic constructions. At the current phase of the project we decided to collect the most frequent constructions only: their absolute frequency had to be more than 250. The dictionary contains roughly 2300 entries and 6500 VPCs. Each construction is illustrated by a corpus example. The verbal entries are presented in alphabetical order primarily. Different kinds of indices are also included in the printed version. The users of this dictionary envisaged to be mainly linguists, working on Hungarian grammars, lexicographers working on bilingual dictionaries and last but not least: advanced level learners of Hungarian, who want to expand their knowledge on the Hungarian nominal verbal collocation relationships. The dictionary is planned to be published both in printed and electronic format. Parts of the algorithm used for this project could be applied to produce other dictionaries, all the more so, as some of them are actually language independent. It is also highly cost effective: both the programming and the lexicographic work required one person year each.

1. Introduction

The significance of using semi-automatic methods for dictionary making has been emphasized by prominent scholars. After the provocative talk of Grefenstette (1998), Rundell (2009) also suggested that the routine tasks of the lexicographers should be gradually replaced by sophisticated programs, so that the lexicographers can concentrate on the crucial tasks (identifying the different senses, writing definitions etc.), which can only be made by trained experts.

Corpus collection and analysis both for diachronic (http://www.nytud.hu/hhc) and synchronic Hungarian texts (http://corpus.nytud.hu/mnsz) has been under way for quite a while (Pajzs 1991, 1997, Oravecz 2002, Váradi 2002). The first volumes of the diachronic corpus based Academic Dictionary of Hungarian has been published (Ittzés 2006). As opposed to the corpus based approach, in the current project we decided to use a clearly corpus driven method (Tognini-Bonelli 2001) for producing the Frequency Dictionary of Hungarian Verb Constructions (FDVC). The only source of this dictionary is the synchronic corpus, which is analyzed by a series of programs (Sass 2009a). Its output is a set of XML format draft entries, which were then hand validated and edited by lexicographers. The dictionary contains the most frequent Hungarian verbs along with their most typical syntactic constructions: 'form and meaning pairings' (Goldberg 2006). Constructions are basically syntactic patterns and the most frequent collocates matching the given pattern. At the current phase of the project we decided to collect the most frequent constructions only: their absolute frequency had to be more than 250. The dictionary contains roughly 2300 entries and 6500 VPCs. Each construction is illustrated by a corpus example. The verbal entries are presented in alphabetical order primarily. Different kinds of indices are also included in the printed version. The users of this dictionary envisaged to be mainly linguists, working on Hungarian grammars, lexicographers working on bilingual dictionaries and last but not least: advanced level learners of Hungarian, who want to expand their knowledge on the Hungarian nominal verbal collocation relationships.

2. Language technology methods applied in the project

The representative corpus of current Hungarian - at the turn of the millennium - contains 187,6 million running words from five different language registers and also includes language variants which are used outside Hungary. The corpus is available for research through the internet: corpus.nytud.hu/mnsz (Váradi 2002). The corpus was POS tagged and disambiguated, the precision rate is 97,5% (Oravecz-Dienes 2002). The POS tagged corpus was first separated into supposed clauses and then these were syntactically analyzed by shallow parsing. The aim of this was not a complete analysis, only a partial syntactic analysis of the verbs and their nominal complements. The typical syntactic patterns and the frequent collocates matching these patterns were collected and arranged with a specialized algorithm which has been described in more detail in (Sass 2009a). The result of the process is a set of files, each containing a draft entry of the dictionary in XML format. The entries are hand validated and edited by the lexicographers in the Xmetal XML editor.

The draft entries contain a set of possible illustrative quotations (usually 10) for each construction. The lexicographers try to select one of these. If there is one they find satisfactory, they only have to click at the 'selected' argument of the citation. If none of the quotations is considered good enough, task specific software can be used to retrieve the verb and its collocates with the specified endings, and select a quotation from the whole corpus. With this tool (http://corpus/nytud.hu/mazsola) (Sass 2008) verb constructions can be efficiently retrieved: the user can ask for a given verb and the nominal with specified suffixes occurring in its contexts. The nominal collocates with the retrieved suffix are presented in decreasing order of salience (Kilgariff -Tugwell, 2001). Each corpus example of the collocates can be listed. It is also possible to search for typical phrases containing at least one verb, either if they are idiomatic expressions or if they are only frequent co-occurrences. (E.g.: the phrase *mosolyt fakaszt* 'make someone smile' can either be retrieved by searching the verb *fakaszt* 'cause, bring forth' co-occuring with a noun with the suffix *–t* 'accusative' or searching the verb *fakaszt* 'cause, bring forth' and the noun *mosoly* 'smile'.) Multi-element constructions can also be retrieved: at the moment up to three different suffixes and/or co-occurring lemmas can be searched and an additional optional running word can also be specified before the search. With this tool highly complex phrases can be efficiently retrieved from the corpus. In many ways this software can be considered as a realisation of the suggestions made in (Pajzs 2002), although it was developed independently.

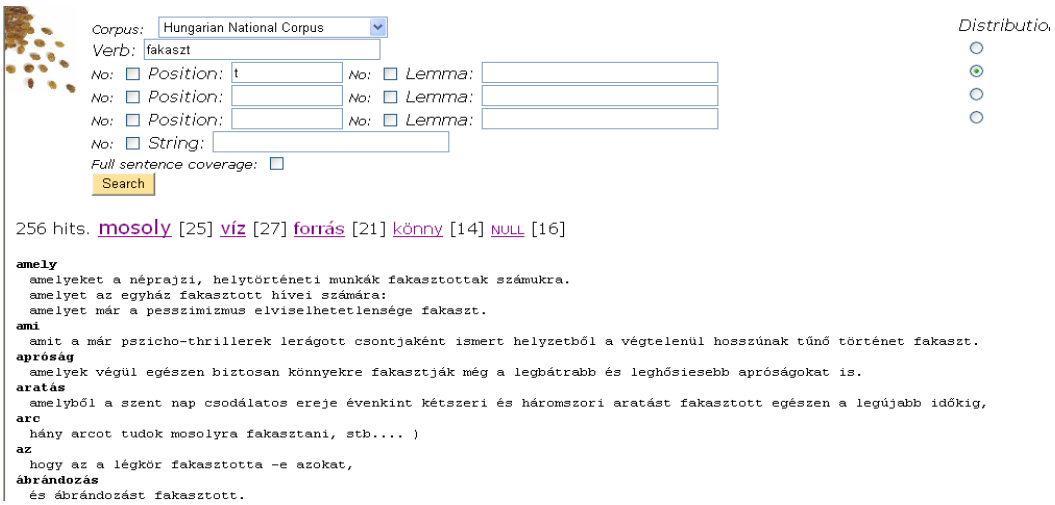

Figure 1. Retrieval of the verb fakaszt 'make someone smile' co-occuring with nouns with the suffix *–t* 'accusative'

Additional citations can be included (by copy and paste plus selection) into the entries of FDVC among the results of Mazsola software, described above. It is also possible to use the default retrieval tool of the corpus, which is available when entering the Mnsz homepage. For our task at hand, the verb argument retrieval tool Mazsola was usually more convenient. In some cases, however, it was safer to use the default retrieval tool, mostly in the case of highly frequent verbs with too frequent collocates.

After the lexicographers have checked, corrected and edited the draft entries, the complete dictionary is generated by another set of programs. The first part of the dictionary contains the entries in alphabetical order, the next part contains the verbs in decreasing order of frequency, and some additional indices are also generated. Among the indices, probably the most interesting for several users can be the alphabetized list of nominal collocates, and the verbs occurring typically in their context.

3. Lexicographic tasks

3.1. The content and structure of the draft entries

Each entry contains the headword verbal lemma and its absolute frequency. This is followed by the most frequent VPCs of the same verb: eg: V+OBJ, V+OBJ+DAT, etc. When a specific noun occurs with a frequency greater than 250 with the given suffix within the same verb phrase, it is also shown in a separate VPC. In the XML version of the sample entry each VPC is indicated by a <pattern> tag, its frequency is included as the value of the attribute freq. The combination of typical suffixes and/or lemmas are called <frame>s. Patterns can include other patterns, when a frequent lemma occurs within a construction which was already specified.

For example the frame of the verb *köt*:

<frame><p c='**-hOz**' l='**feltétel**'/><p c='**-t**'/></frame>:
köt ***feltétel*** + ***hOz*** +***t***
'bound' 'condition'+Allativus +OBJ 'subject sth to conditions'
is a specialised subcase of the frame <frame><p c='**-hOz**'/><p c='**-t**'/></frame>:
köt +**hOz** +**t**
'bound' Allativus OBJ.

Part of a sample entry can be seen in Figure 2, with some editorial comments in English.

```
  <?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE fdvc SYSTEM 'fdvc.dtd'>
<?xml-stylesheet type='text/xsl' href='fdvc_plain.xsl.xml'?>
<fdvc>
<entry remark='OK'>
<verb lemma='köt' freq='23634'/>
<pattern freq='1308'> <frame><p c='-hOz'/><p c='-t'/></frame>
<type str='2:02' len='2' fixed='0' free='2'/>
<cits>
  <cit type='sentence'>A szocialisták kétharmados országgyűlési többséghez kötnék az ország területéről induló
harci cselekmények engedélyezését.</cit>
  <cit type='sentence'>hanem a bejelentés hatósági publikálásához köti.</cit>
  …
  <cit type='sentence' selected='yes'>Kicsit följebb a rongyot kötötte hozzá a zsineghez.</cit>
  ...
</cits>
<pattern freq='387' idiom='yes'>
<frame><p c='-hOz' l='feltétel'/><p c='-t'/></frame>
<type str='3:11' len='3' fixed='1' free='1'/>                    Comments on the example sentences
```

```
<cits>
 <cit type='sentence'>vagy feltételhez köti.</cit>                    No explicit object
 <cit type='sentence'>illetve feltételhez kötheti.</cit>              No explicit object, the verb is in derivative form
 <cit type='sentence'>illetve feltételekhez kötheti.</cit>            No explicit object, the verb is in derivative form
 <cit>feltételhez kötheti,</cit>                                      No explicit object, the verb is in derivative form
 <cit>viszont olyan feltételekhez köti azt,</cit>                     The object is the pronoun az
 <cit type='sentence'>viszont a részvételt szigorú feltételekhez kötnék.</cit>   Surplus complement
 <cit>vagy többoldalú megállapodások megkötését ahhoz a feltételhez köthetik,</cit>
                                                                      The verb is in derivative form, surplus
                                                                      complement
 <cit type='sentence'>vagy mérőeszközhasználatot újabb feltételekhez kötheti.</cit>
                                                                      The verb is in derivative form
 <cit>vagy feltételhez köti,</cit>                                    No explicit object
 <cit type='sentence'>vagy feltételhez kötheti.</cit>                 No explicit object, the verb is in derivative form
 <cit type='sentence'>vagy feltételekhez kötheti.</cit>               No explicit object, the verb is in derivative form
 <cit>vagy ezek végzését feltételhez kötheti;</cit>                   The verb is in derivative form
 <cit>ugyanis szigorú feltételekhez kötik,</cit>                      No explicit object, surplus complement
 <cit>több feltételhez kötötték a repülést,</cit>                     Good candidate
 <cit type='sentence'>Szűkebb feltételekhez kellene kötni a kényszerítés lehetőségét.</cit>
                                                                      The verb is in derivative form
 <cit>Szigorú feltételekhez kötötték</cit>                            No explicit object
 <cit selected='yes'>Szeretetét feltételekhez köti,</cit>             Good example, selected by the lexicographer
 <cit type='sentence'>s feltételekhez kötné a költségvetés megszavazását.</cit> Correct candidate
 <cit>s ezt ahhoz a feltételhez kötötte,</cit>                        The object is the pronoun ez,
 <cit>s ez szigorú feltételekhez köti a tartásukat,</cit>             Good candidate
</cits>
</pattern>
</pattern>

.....
</entry>
</fdvc>
```

Figure 2. Part of the sample entry in XML format

3.2. Editing the entries

The draft entries are edited in the XMetal editor. The lexicographer decides whether the automatically created VPCs are correct. The overall precision rate was over 94%. It is important to emphasize that the editors only mark constructions to delete, which are clearly erroneous. The errors can be caused by the application of the set of programs, either in the phase of the morphological analysis or during the disambiguation process. A typical error is caused by some ambiguous verbal suffixes: e.g. the same verbal suffix can mean that the verbal predicate has a definite object, but it can also refer to intransitive usage. For example, the suffixed verb form *bámultam* can mean 'I was staring at him' or 'I was staring'. To illustrate the correct construction clearly, the editors tried to find example sentences with explicit [1] objects: *Zimonyi a cipőjét bámulta* 'Zimonyi (surname) was staring at his shoes'.

Because of the ambiguous meaning of this suffix, the program sometimes suggested that some intransitive verbs had transitive uses as well. The editors had to decide carefully, if these kinds of constructions were really incorrect, because in some cases a new meaning of a verb had appeared which was not yet registered by the existing dictionaries. Sometimes the program suggested that a clearly transitive verb had intransitive uses as well, again, a number of examples had to be checked to decide, if they were simply errors or real examples of new

[1] In Hungarian, the subject and object can be expressed by verbal suffixes. E.g. the running word *szeretlek* means 'I love you', which is a special construction in which the root is *szeret* and the suffix *lek* expresses that the subject of the predicate is in the first person singular and its object is in the second person singular. By 'explicit' subject and object we mean the ones, which are not expressed solely by verbal suffix.

usage. The errors made by the programme were usually caused by the inefficiency of the earlier disambiguation process. In some cases either the whole entry or some constructions of the entry were marked to be moved into another entry, again as a result of improper disambiguation.

	Original	Marked to Delete	Marked to Move	Total	Precision rate %
Number of draft entries	2338	89	43	2206	94,3
Number of constructions	6853	340	32	6481	94,6

Figure 3. Number of entries and constructions

Some (usually 10) example sentence candidates were offered by the programme. When choosing the example sentences, we were encouraged by the results of (Kilgariff et al, 2008). In our case, the lexicographers could usually select and mark one of the candidates as an appropriate illustrative quotation. Although half of the corpus was based on journals and periodicals, the editors intended to show the variety of the corpus as much as possible, several examples were selected from novels, short stories, tales, chat rooms or scientific texts. Sentences with (either politically or otherwise) provocative or possible offending meanings were avoided. The grammatical form of the chosen illustration is even more important. The grammatical selection criteria were:

Choose sentences which are 'complete', namely, they should have an explicit object, when the frame contains an object, and possibly an explicit subject as well.
The predicate should not be in derivative form.
There should not be other complements in the sentence than those which are indicated in the pattern.
If possible, the suffixes which are to be illustrated should appear on content words, not only on pronouns.

The program offers too many incorrect candidates at the moment, because it is a general purpose corpus query tool not specially tailored to find good dictionary examples. If a more detailed version of the current dictionary is prepared later, the program can be further improved by some additional automatic selection criteria. In the current version, when none of the 10 candidates were considered good enough, the lexicographer retrieved each sentence containing the given VPC by the on-line retrieval tool Mazsola, which was described in section 2. Sometimes this phase makes clear that most of the examples are invalid, in which case the whole VPC is marked to be deleted.

3.3. Results from the editor's viewpoint

köt [23634]
'bind', 'tide','attach','fasten'

köt -hOz -t[1308]	*a rongyot kötötte hozzá a zsineghez*
'fasten to + OBJ '	*'He fastened the rug to the string.'*
köt feltétel-hOz –t [387]	*Szeretetét feltételekhez köti,*
'subject to condition + *OBJ'*	*'his love is conditional'*
köt -t[1054]	nyakkendőt köt
'tie+OBJ'	*'he puts on his tie'*
köt szerződés-t [747]	*öt lemezre szóló szerződést kötöttünk,*
'make a contract'	*'we made a contract for five discs'*
köt megállapodás-t [284]	*Tisztességes megállapodást kötöttünk.*

'make an agreement'	*'we have made an honest agreement'*
köt szerződés-t –vAl [970]	*Valamennyi alapítvánnyal pontos szerződést kötnek,*
'make a contract + with sb'	*'they make a correct contract with each of the foundations'*
köt megállapodás-t –vAl [476]	*új megállapodást köt a kerületi önkormányzattal*
'make an agreement with sb',	*'he makes a new agreement with the municipiality'*
köt –hOz [419]	*Köt még valami a szülőföldedhez?*
'attach to'	*'Are you still attached to your homeland?'*
köt -t -vAl[331]	*nem köt elvtelen kompromisszumokat a hatalommal.*
'make an agreement/compromise with'	*'he does not make an unprincipled compromise with the authorities'*
köt -rA -t[291]	*Madzagot kötöttem lábfejemre,*
'fasten to+OBJ'	*'I fastened a string to my foot'*

Figure 4. The sample entry *köt*

As this is a frequency dictionary, it does not contain definitions or non-Hungarian equivalents. However, with the structure made automatically out of the patterns the entries suggest some kind of sense distinction. While the general pattern *köt+t* can be used in any of the several senses of this word, the collocations *szerződés* 'contract', *megállapodás* 'agreement' make the actual sense unambiguous. We completely agree with John Sinclair's claim: language is actually built of semi-pre-constructed phrases, rather than words. He even states: 'Several long accepted conventions in lexicography were called into question - for example the idea that a word could inherently have one or more meanings. The working assumption was that when these meanings were explicated (or translated, in a bilingual dictionary) and, in the better dictionaries, exemplified, the lexicographer's job was done. This practice proved incapable of organising the strong, recurrent patterns that were shown by corpus analysis to be present in the way words were used in texts; the importance of the surrounding language far outweighed the question of how many meanings and how they were related to each other.' (Sinclair 1998: 2.) He also states: 'many, if not most, meanings require the presence of more than one word for their normal realisation'. He concludes that the word is not the best starting point for a description of meaning because meaning arises from words in particular combinations.

Since we also believe that the multiword lexical items should be the core of the dictionaries of the future, we simply presented each multiword phrase, which was found by the programme and considered correct by the editors. No selection was made according to other criteria (i.e. if it was an idiom or a frequent collocation only). Several kinds of multiword expressions can be found in it:

- phrasal verbs, e.g.: *részt vesz* 'take part',
- idioms, e.g.: *munkához lát* 'get down to work'
- function verb phrases: e.g: *tanácsot ad* 'give advice',
- frequent collocations, eg.: *bánja tettét* 'regret his action'

The over two thousand multiword units (i.e.:2041) presented both under their verbal entry and at their nominal part in the index, is the main merit of this dictionary. Although several dictionary of idioms and word phrases were published in the last decade (Bárdosi 2003, Forgács 2003, T. Litovkina 2005), they mainly concentrate on phrases with special or strange meanings. Since our dictionary only takes into account the frequency of the constructions, it is hoped to be a valuable resource for further lexicographic and grammatical studies. Some of the frequent multiword units should be treated as separate lexical items, and handled accordingly, with proper equivalents in bilingual dictionaries. Some of these can only serve as suggestions for sample sentences in future dictionaries/grammars.

From the indices one can also check, which verbs are used mostly with the nouns. In Figures 5 and 6 two examples illustrate this.

	Verb	Equivalent	Frequency
ember 'human'	van	exist (present+past)	7442
	lesz	exist (future)	1763
	mond	tell	1001
	tud	know	971
	lát	see	961
	él	live	939
	meghal	die	781
	érez	feel	631
	szeret	love	588
	hisz	believe	586
	gondol	think	555
	ismer	know	546
	tesz	make	371
	kap	receive	349
	néz	watch	311
	csinál	make	286
	ad	give	281
	találkozik	meet	280
	vár	wait	268
	megöl	murder	260
	vesz	take	260

Figure 5. Verbal collocates of the noun *ember* 'human', ordered according to frequency

Beside describing the collocational relationships, our entries can also help to give a sketch of the word, similar to Kilgarriff's (2001). For example, the noun *kapcsolat* 'relationship' is used in the following phrases:

kapcsolatban áll vkivel	'be in touch with sb'
felveszi a kapcsolatot	'contact sb'
fenntartja a kapcsolatot	'maintain relations'
kapcsolatba kerül vkivel	'get into touch with sb'
kapcsolatba lép	'contact sb'
kapcsolatot tart	'maintain relations'
kapcsolatot teremt	'establish relations'
kapcsolata van vkivel	'have a relationship with sb'
kapcsolatban van vkivel	'be in touch with sb'

Figure 6. The verbal collocates of the noun *kapcsolat* 'relationship'

The inclusion of carefully selected corpus examples adds a further value to this simple frequency dictionary. When selecting the examples, we had to realise, that we were in agreement with Hanks's (2005) observation: the so called metaphorical or figurative meanings of the words are much more frequent than the literate meanings. When the editors selected examples for the general form of the pattern (*köt+hOz+t*), they tried to find sentences

in which the verb was used literally (‘fasten’ in the case of *köt*), but this was usually only possible by applying some extra ‘tricks’ (i.e. searching certain nominal collocates, by the help of the on-line retrieval tool).

The corpus driven presentation of the most frequent syntactic constructions of the verbs is a novelty in Hungarian lexicography. One of the indices presents the content ordered according to the constructions, so that linguists can see which verbs share similar syntactic patterns. A sample is presented in Figure 7.

elmagyaráz -nAk –t	‘explain sth to sb’
elmesél -nAk –t	‘tell sth (a story) to sb’
elmond -nAk –t	‘tell sth to sb’
elnevez -nAk –t	‘name sth’
elnéz -nAk –t	‘forgive sb’
engedélyez -nAk –t	‘allow sth to sb’
érez -nAk –t	‘feel sth’
értékel -nAk –t	‘appreciate sth’
felajánl -nAk –t	‘offer sth to sb’
felel -nAk –t	‘aswer sb’
felfog -nAk –t	‘undestand sth’
felró -nAk –t	‘blame sb for sth’
gondol -nAk –t	‘think that sb is sth’
hisz -nAk –t	‘believe that sb is sth’
hív -nAk –t	‘name sb’
hoz -nAk –t	‘bring sth to sb’
ígér -nAk –t	‘promise sth to sb’
ír -nAk –t	‘write sth to sb’
ismer -nAk –t	‘know sb as sb’
ítél -nAk –t	‘consider sth as’
javasol -nAk –t	‘suggest sth to sb’
jelent -nAk –t	‘mean sth to sb’
juttat -nAk –t	‘allocate sth to sb’
kap -nAk –t	’get sth as sth’
kér -nAk –t	’ask sth for sb’
készít -nAk –t	’prepare sth to sb’

Figure 7. Part of the list of verbs, occuring in the frame „-nAk –t’ ‘DAT+OBJ’

These lists can be essential for creating better dictionaries and grammars. Most of these phrases should be present in bilingual dictionaries, because they hardly ever have a word to word, suffix to preposition translation.

4. Conclusion

We hope that this frequency dictionary is a valuable resource for further researches. Beyond publishing the dictionary itself, we also wished to show, that the current NLP tools can greatly help the traditional lexicographers by analysing and presenting the corpus data according to the needs of the task at hand.

We are aware that our attempt is just a small step towards automatic dictionary creating, mainly because our dictionary is a ‘meaningless dictionary’ (Janssen 2008), in the sense that it

does not contain sense distinctions and definitions or second language equivalents. We are convinced, however, that parts of the algorithm used for this project could be applied to produce other dictionaries, all the more so, as some of them are actually language independent. They were already tested on Danish texts (Sass 2009b).

It is also important to emphasize that this project was highly cost effective: both the programming and the lexicographic work required one person year each. The actual editorial work lasted 5 months with a full time and a part time lexicographer. A few months were also spent on experimenting before the actual start of the editing, and some for finalizing the complete work.

Our very next project will be an adjective–noun collocation dictionary of Hungarian. We also foresee future projects, where we attempt to step towards some kind of automatic semantic recognition as well, by connecting our tools with the Hungarian Wordnet database.

References

Bárdosi, V. (2003). *Magyar szólástár. Szólások, helyzetmondatok, közmondások értelmező és fogalomköri szótára.* Budapest: Tinta Könyvkiadó.

Forgács, T. (2003). *Magyar szólások és közmondások tára*. Budapest: Tinta Könyvkiadó.

Goldberg, A. E. (2006). *Constructions at Work.* Oxford: Oxford University Press

Grefenstette, G. (1998). 'The future of linguistics and lexicographers: Will there be lexicographers in the year 3000?' In *Proceedings of EURALEX 98*, Liège: University of Liège. 25–41

Hanks, P. (2005). 'Metaphors and meanings: a lexicographical approach to corpus analysis'. In Kiefer F.; Kiss G.; Pajzs J. (ed.). *Papers in Computational Lexicography, COMPLEX 2005*. Budapest: Research Institute for Linguistics, HAS. 81–106

Janssen, M. (2008). 'Meaningless dictionaries'. In Bernal, E.; DeCesaris J. (eds.). *Proceedings of the XIII EURALEX International Congress*, Barcelona: Institut Universitari de Lingüistica Aplicada, Universitet Pompeu Fabra. 409–420.

Ittzés, N. (2006). *A magyar nyelv nagyszótára* I.II. Budapest, Akadémiai Kiadó.

Kilgarriff, A.; Tugwell, D. (2001). 'Word Sketch: Extraction and display of significant collocations for lexicography'. In *Proceedings of the 39th Meeting of the Association for Computational Linguistics, workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation*, Toulouse: Association for Computational Linguistics. 32–38

Kilgarriff, A.; Husák, M.; McAdam, K.; Rundell, M.; Rychly, P. (2008). 'GDEX: Automatically finding good dictionary examples'. In Bernal, E.; DeCesaris J. (eds.). *Proceedings of the XIII EURALEX International Congress*, Barcelona: Institut Universitari de Lingüistica Aplicada, Universitet Pompeu Fabra. 425–432.

Oravecz, Cs. (2002). 'Large scale morphosyntactic annotation of the Hungarian National Corpus'. In Hollósi B.; Kiss-Gulyás J. (ed.). *Studies in Linguistics*, Volume VI. Debrecen: Institute of English and American Studies, University of Debrecen. 277–298.

Pajzs, J. (1991). 'The Use of a Lemmatized Corpus for Compiling the Dictionary of Hungarian'. In *Using Corpora Proceedings of the 7th Annual Conference of the OUP & Centre for the New OED and Text Research.* Waterloo: University of Waterloo Centre for the New OED. 129-136.

Pajzs, J. (1997). 'Synthesis of results about analysis of corpora in Hungarian'. In *Linguisticae Investigationes* XXI-2. Amsterdam, John Benjamins. 349-365.

Pajzs J. (2002). 'A corpus based investigation of collocations in Hungarian.' *Proceedings of EURALEX 2002*. Copenhagen: University of Copenhagen. 831–840.

Rundell, M. (2009). 'The road to automated lexicography: First banish the drudgery... then the drudges?' In *Proceedings of eLexicography in the 21st Century Conference*, Louvain-la-Neuve, Louvain: Université Catholique de Louvain.

Sass, B. (2008). 'The Verb Argument Browser'. In Sojka, P., Horák, A., Kopecek, I., Pala, K. (eds.). *11th International Conference on Text, Speech and Dialog, TSD,* Proceedings. Lecture Notes in Computer Science: Springer. 187-192.

Sass B. (2009a). 'A unified method for extracting simple and multiword verbs with valence information and application for Hungarian'. In *Proceedings of RANLP 2009, Bukarest: Bulgarian Academy of Sciences*. 399-403.

Sass B. (2009b). 'Verb Argument Browser for Danish'. In *Proceedings of the 17th Nordic Conference of Computational Linguistics, NoDaLiDa 2009*, Odense: University of Southern Denmark. 263–266.

Sinclair, J. (1998). 'The lexical item'. In Weigand E. (ed.). *Contrastive Lexical Semantics*, Amsterdam and Philadelphia: John Benjamins. 1–24.

T. Litovkina A. (2005). *Magyar közmondástár*. Budapest: Tinta Kiadó.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam and Philadelphia: John Benjamins.

Váradi T. (2002). 'The Hungarian National Corpus'. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (LREC2002), Paris: ELRA, 385–389.