

Finding Multiword Term Candidates in Croatian

Marko Tadić* and Krešimir Šojat†

*Department of Linguistics and †Institute of Linguistics
Faculty of Philosophy, University of Zagreb
Ivana Lučića 3
10000 Zagreb, Croatia
{mtadic, ksojat}@ffzg.hr

Abstract

The paper presents the research in the field of statistical processing of a corpus of texts in Croatian with the primary aim of finding statistically significant co-occurrences of n-grams of tokens (digrams¹, trigrams and tetragrams). The collocations found with this method present the list of candidates for multi-word terminological units submitted to terminologists for further processing i.e. manual selecting of the “real terms”. The statistical measure of co-occurrence used is mutual information (MI³) accompanied with linguistic filters: stop-words and POS. The results on non-lemmatized material of a highly inflected language such as Croatian show that MI measure alone is not sufficient to find satisfactory number of multi-word term candidates. In this case the usage of absolute frequency combined with linguistic filtering techniques gives broader list of candidates for real terms.

1. Introduction

The statistical and/or pattern based approach to collocation detection has been known in the corpus processing for some time now. There is a number of works dealing with either statistical approach alone (Church & Hanks 1990; Church et al. 1991; Smadja 1993; Su et al. 1994), in combination with linguistic filtering (Daille 1994 and 1995; Hatzivassiloglou 1994; Tiedemann 2000; Jacquemin 2001) or using linguistic patterns primarily (Kilgariff & Tugwell 2002). The common characteristics of these approaches is that they were tested mostly on typologically similar languages (English, French etc.) with almost insignificant inflectional variation i.e. small amount of different word-forms for a lemma. Unlike in (Vintar 2000 and 2002) where the detection of “termhood” of the Slovenian multi-word units (MWU) intended for translation equivalents determination was the primary objective, we wanted to test the purely statistical

approach to find out collocation candidates in similarly highly inflected language in order to get at least the basic statistics about the expectation of n-grams in Croatian.

The concrete task being presented in this paper is a part of a larger project of collocation detection in Croatian texts. Collocations retrieved within the project should serve as term candidates for building the domain-specific terminological glossaries.

In the Section 2 the aim of the project is described and the position of the research presented here is explained. Section 3 describes the corpus and gives its parameters while the Section 4 describes its processing. Section 5 explains method more precisely and gives the discussion of results. The paper ends with conclusions and suggestions for further research directions.

2. Aim of the research

In October 2002 the Ministry of European Integration of the Republic of Croatia has started the pilot joint-project with the Institute of Linguistics named *The compilation of glossary and harmonization of terminology in the field of banking, insurance and other financial services*. The basic idea of the project is to take a financial part (ca 2 Mw) of *Acquis communautaire* in English and build a glossary which would serve the translators as obligatory terminological resource in the process of translating EU legislative documents from that field into Croatian. The glossary is being built by terminologists who are selecting the “real” terms from the list of MWU (digrams, trigrams and tetragrams) obtained by statistical methods but it will be additionally filled with terms from other sources such as other glossaries constructed in traditional way — by hand. Once established, this glossary will serve as the resource for term marking in original texts which proved itself to be indispensable tool for translators. In such a way all the “official” terms are already marked and signaled in the original texts.

The analogous task of building the list of MWU term candidates has been performed on Croatian texts, which were collected in a corpus of approximately 500 Kw. The basic assumption was that some characteristic MWU in Croatian can be detected and used as term candidates or translational equivalents on the left side of a bilingual glossary. Secondly, Croatian MWUs will be used in the process of translating Croatian legislation to English, which is also, one of tasks set by the same Ministry. This

¹ We use the term 'digrams' instead of more common 'bigrams' because it is an element of proper series of terms of Greek origin: di-, tri-, tetra-, penta-, hexa-, etc. -grams. 'Bi-' is prefix of Latin origin which combines with Greek root forming Latin-Greek hybrid.

paper deals only with detection of digrams, trigrams and tetragrams in Croatian.

3. Corpus statistics

The corpus consists of 186 documents from the field of finances. The documents comprise recommendations (109 documents), decisions (35 documents) by the Croatian National Bank and law texts (42 documents) dealing with different aspects of financial sector. The sizes of documents vary from 82 to 48,018 tokens with average size of 2812.22 tokens. The document structure is mostly uniform with typical structure of legal document separated into firmly delimited sections and/or articles.

The corpus has 509,012 tokens including punctuation and tags. When these has been stripped off, the amount of 460,664 tokens was reached (including digits).

Documents included in the corpus are being issued from 1993 until today so we are dealing with contemporary data.

The basic corpus processing consisted of several standard steps. The first one was text conversion from DOC/RTF format to XML. After sentence segmentation, the next step was tokenization. For all tasks we used our own tools (2XML and Tokenizer) developed during the building of Croatian National Corpus (HNK²) (see in Tadić 2002). For further processing the tokenized (or verticalized) corpus was imported into a database.

The corpus was not POS-tagged or lemmatized because the POS-tagger for Croatian, which would enable the automatic tagging and further use of tagged material, is not developed yet.

4. Processing

4.1. Frequency and type/token ratio

The statistical processing was organized on several levels. The first one was the processing of individual words i.e. getting the list of types accompanied by their frequency. In the corpus of 460,664 tokens and 24,286 types were found what gives a 1:18.97 type:token ratio. The ratio is unexpectedly high having in mind the data from *Croatian Frequency Dictionary* (Moguš et al. 1999) where it is 1:8.38 or (Allen & Hogan 1998) where it is around 1:11.8. Possible explanation for this is that the vocabulary in the field of finance shows less variation in inflection as well as limited number of different lexical entries than the general vocabulary.

4.2. Detection of co-occurrence

The second level of processing was to establish word co-occurrences. We limited the scope of our research on digrams, trigrams and tetragrams. The statistical measure

of co-occurrence used in this research was mutual information in its variation called cubic association ratio or MI³ (McEnery et al. 1997:222; Oakes 1998:172).

The digrams (trigrams and tetragrams) are defined as sequences of two (three or four) tokens belonging to the same sentence and uninterrupted by punctuation. There are two reasons for so strict definition: 1) MWU terms are expected to be used generally as a whole without much interruption by stop-words; 2) punctuation is not expected within the MWU term because it usually introduces a break and/or new (sub-)unit. The possible case of abbreviations is not relevant since in Croatian the abbreviations are written with capitals and without punctuation (e.g. *EU*, *PEN*, *UN* and not *E.U.*, *P.E.N.*, *U.N.*). If the abbreviation changes case, it is written as *EU-a*, *PEN-u*, *UN-om* but the tokenizer takes care about those cases and treats them as compounds.

Although it has been already shown that lemmatization of Croatian texts gives better statistical results in collocation detection (Tadić et al. 2003), in this work we performed all calculations on non-lemmatized material in order to see whether is it possible to avoid the time-consuming and painstaking process of lemmatization which also includes morphosyntactic description (MSD) and sometimes POS disambiguation, as well. Since in Croatian the “internal” homography (covering different word-forms of the same lemma e.g. different cases of the same noun but with the same form) is far more frequent than “external” homography (covering word-forms of the same form belonging to two or more different lemmas with same or different POSs), and since no MSD information will be used in this research apart from POS, the starting assumption was that with avoidance of lemmatization step, much of valuable information would not be lost. The lists of co-occurrences for digrams, trigrams and tetragrams were produced and stored in the database with their frequencies and MI values calculated.

4.3. Filtering

Additionally, like (Smadja 1993) and unlike (Daille 1995), linguistic filtering in two steps was performed after statistical processing. The first one was filtering out stop-words. In order to do that, a stop-word list of 954 prepositions, pronouns, adverbs, particles and numerals was built and applied to the already calculated MWUs. Combinations containing stopwords were then filtered out.

The second filter was POS information, which was mapped from the list of word-forms, generated by *Croatian Morphological Lexicon* (Tadić 1994; Tadić & Fulgosi 2003). In the cases of external homography manual disambiguation was performed. Only five POS tags were considered in term-candidates and these are N for nouns, A for adjectives, V for verbs, M for numerals and Y for abbreviations. Since the *Croatian Morphological Lexicon* is MULTEXT-East v2.1 conformant (Erjavec et al. 2003), so are the generated MSD and POS tags.

At the end of processing the manual evaluation of detected MWUs was done in order to establish the ratio between statistically detected MWUs and “real” terms.

² We suggest this acronym for the Croatian National Corpus (Hrvatski nacionalni korpus) since the CNC is already used for Czech National Corpus and HNC appears to be used for the Hungarian National Corpus.

5. Method and results discussion

Although the statistical processing was done on the whole corpus, the manual inspection of real terms was limited to MWUs with frequency higher than 4. Having in mind the inflectional complexity of Croatian which can, and usually does, produce a lot of single occurrences of word-forms even in very large corpora, we set the threshold above the frequency of 4 in order to keep the amount of data for manual inspection within the processable limits. This threshold proved to be enough to show the difference between the candidates selected by MI and by frequency accompanied by linguistic filters.

5.1. Digrams

The results for digrams are given in the Table 1.

2gram tokens	390,102
2gram types	137,535
2grams:MI3>0	1,140
2grams:MI3>0,real terms	499
2grams:frq>4	13,103
2grams:no stop-words	68,445
2grams:no stop-words,both POS	68,222
2grams:no stop-words,both POS,real tr.	2,114

Table 1: Statistics for digrams

The results show that of 137,535 different digrams only 1,140 (0.83%) have MI higher than 0 out of which only 499 (39.39%) are evaluated as real terms. When digrams were ordered by descending frequency and after application of both linguistic filters and manual evaluation of 13,103 (9.53%) digrams with frequency higher than 4, the real terms were detected in 2,114 cases (16.13%). Although the number of real terms found from the list of candidates is by percentage smaller (39.39% against 16.13% in the favor of MI3) it clearly outperforms the MI because the MI calculated on non-lemmatized text gives more that four times less real terms (499 against 2114).

The characteristic digram POS sequences for real terms with accompanied frequency are:

A+N	1305
N+N	737
V+N	32
N+Y	28
N+A	9
A+Y	2
M+A	1

Table 2: Statistics of digram POS sequences

It shows intuitively predicted patterns but such predominance of A+N combination was not expected.

5.2. Trigrams

While the application of MI for two occurring words is quite common, the usage of MI for calculating trigram co-occurrences is quite rare. There are only few papers dealing with this matter and each of them approaching it from different angles.

While Su et al. (1994) defines the MI of trigram as:

$$I(x; y; z) \equiv \log_2 \frac{P(x, y, z)}{P(x) \times P(y) \times P(z) + P(x) \times P(y, z) + P(x, y) \times P(z)}$$

Boulis (2002) used different approach. One of his suggestions was to calculate average intra-cluster pairwise MI, which is calculated as:

$$I(x; y; z) \equiv \frac{I(x, y) + I(x, z) + I(y, z)}{3}$$

This second approach was more useful in our case since we could use already calculated MI values for direct digrams $I(x, y)$ and $I(y, z)$ and only MI for indirect digram $I(x, z)$ had to be calculated. We used the MI³ variant for calculation of each pair and called this whole calculus MI3a3.

In addition we have used another formula combining also MI³ calculations for pairs of initial/final digram with ending/starting single element. We called this calculus MI3b3:

$$I(x; y; z) \equiv \frac{I(xy, z) + I(x, yz)}{2}$$

The results for trigrams are given in the Table 3:

3gram tokens	333,783
3gram types	211,225
3grams:MI3a3>0	362
3grams:MI3b3>0	861
3grams:MI3a3>0,real terms	114
3grams:MI3b3>0,real terms	103
3grams:frq>4	6,781
3grams:no stop-words	1,372
3grams:no stop-words,all POS	1,362
3grams:no stop-words,all POS,real tr.	551

Table 3: Statistics for trigrams

The results show that of 211,225 different trigrams only 362 (0.17%) calculated with MI3a3 and 861 (0.41%) calculated with MI3b3 have MI higher than 0. The real terms were recognized in 114 (31.49%) and 103 (11.96%) cases respectively. The application of linguistic filters in combination with frequency yielded 6,781 (3.21%) trigrams with frequency higher than 4 and 551 real terms (8.13%) out of them. The frequency with linguistic filtering gives again more real terms (114/103 against 551).

The characteristic trigram POS sequences with accompanied frequency (>1) for real terms are:

N+A+N	213
A+A+N	138
A+N+N	84
N+N+N	77
A+N+Y	15
N+N+Y	6
V+A+N	5
N+A+A	3
N+N+A	2

Table 4: Statistics of trigram POS sequences

The characteristic POS sequences for digrams and trigrams show almost the same structure and distribution as in (Loukachevitch and Dobrov 2003:60) who were semi-automatically constructing the Thesaurus of Russian official documents. This could lead to conclusion that similar methods applied to other Slavic languages would probably yield similar results.

5.3. Tetragrams

The same problem of finding the appropriate formula for calculating trigrams with MI occurred in the case of tetragrams. Furthermore, with the number of tokens included the possible combinations calculating individual MI pairs is complicated even more.

The formula, which we used for MI3a4, takes into account the average value of all 6 possible pairwise MI³ calculated within a tetragram:

$$I(w; x; y; z) = \frac{I(w, x) + I(w, y) + I(w, z) + I(x, y) + I(x, z) + I(y, z)}{6}$$

As in the case of trigrams we tested another formula in the form of MI3b4 which calculates the relation of starting/ending elements with initial/final trigrams.

$$I(w; x; y; z) = \frac{I(wxy, z) + I(w, xyz)}{2}$$

In such a way the average of MIs between the first trigram and the last element of tetragram on the one side and on the other side between the first element of tetragram and the last trigram of the tetragram is calculated.

The results for tetragrams are given in the Table 5:

4gram tokens	288,702
4gram types	222,521
4grams:MI3a4>0	1,007
4grams:MI3b4>0	1,330
4grams:MI3a4>0, real terms	36
4grams:MI3b4>0, real terms	45
4grams:frq>4	3,139
4grams:no stop-words	354
4grams:no stop-words,all POS	351
4grams:no stop-words,all POS,real tr.	138

Table 5: Statistics for tetragrams

The results show that of 222,521 different trigrams only 1,007 (0.45%) calculated with MI3a4 and 1,330 (0.60%) calculated with MI3b4 have MI higher than 0. The real terms were recognized in 36 (3.58%) and 45 (3.38%) cases respectively. The application of linguistic filters in combination with frequency yielded 3,139 (1.41%) trigrams with frequency higher than 4 and 138 real terms (4.40%) out of them. The frequency with linguistic filtering gives again more real terms (36/45 against 138).

The characteristic tetragram POS sequences with accompanied frequency (>1) for real terms are:

N+A+A+N	33
A+N+N+N	24
A+N+A+N	22
N+N+A+N	15
N+A+N+N	13
A+A+A+N	8
N+N+N+N	7
A+A+N+N	6
A+A+N+Y	2

Table 6: Statistics of tetragram POS sequences

It may be seen in Tables 1, 3 and 5 that the precision (real term coverage) declines with the growth of the number of elements in an n-gram. It also seems that experiments with formulae for different kind of MI calculation does not give the predictable results because of different behavior of trigrams and tetragrams. What could also be

noticed is constant better performance of pure frequency accompanied by two linguistic filters than either MI calculation:

2grams:MI>0, real terms	499
2grams:no stop-words,both POS,real tr.	2,114
3grams:MI3a3>0, real terms	114
3grams:MI3b3>0, real terms	103
3grams:no stop-words,all POS,real tr.	551
4grams:MI3a4>0, real terms	36
4grams:MI3b4>0, real terms	45
4grams:no stop-words,all POS,real tr.	138

Comparing these methods shows that MI gives less candidates but more of them are real terms (17,96% average) while frequency with linguistic filters gives more candidates but less of them are real terms (9,55%). On the other hand in absolute values the latter method gives in average 4.26 times more real terms than the former and clearly outperforms it. Figure 1. gives a sample from the 138 real tetragram terms with clear indication how frequency with linguistic filtering outperforms MI values.

6. Conclusions and Future work

We have shown the method for finding MWU term candidates in Croatian corpus of about 500 Kw. It includes statistical approach, which has primarily used the MI³ for retrieving digrams, trigrams and tetragrams. In the second stage, the frequency accompanied with linguistic filters (stop-words and POS) has been used for the same task yielding better results in absolute numbers. The conclusion that can be drawn is that MI³ alone is not sufficient for selection of real term candidates in non-lemmatized texts. Having in mind the results on lemmatized text (Tadić et al. 2003), one of possible future directions would be to compare the application of MI on lemmatized texts with frequency accompanied by linguistic filters to see which method gives better results.

Furthermore similar POS sequences were found for Russian digrams and trigrams in Loukachevitch and Dobrov (2003). Since Croatian and Russian are typologically and genetically close languages the similar results could be expected for other Slavic languages.

Including MSD beside the POS data would give an insight into the characteristic MSD sequences such as Nn+Ng, Nn+Ag+Ng or Nn+Ag+Ag+Ng where *n* and *g* denote nominative and genitive cases. This could be achieved once the POS-tagger and (semi)automatic lemmatizer for Croatian will be fully developed.

Other statistical measures for co-occurrence and possible "termhood" detection of MWUs (e.g. LogL, as in Vintar (2002) which was applied to Slovenian, or Dice-coefficient or z-score) will be investigated in detail in the future work on both lemmatized and non-lemmatized texts.

References:

- (Allen & Hogan 1998) J. Allen & C. Hogan, 1998. *Expanding Lexical Coverage of Parallel Corpora*, In Proceedings of LREC1998, Granada, Spain, ELRA, 1998, pp. 747-754.

4GR_RAZLICNICA_L2 : Table													
	P1234	FRQ1234	MI4	MI3	EU	SW1	SW2	SW3	SW4	POS1	POS2	POS3	POS4
▶ guverner hrvatske narodne banke	58	3,825980287338	-0,779275777	+	n	n	n	n	N	A	A	N	
konsolidiranog proračuna središnje države	56	5,646128040754	-1,118857873	+	n	n	n	n	A	N	A	N	
savjet hrvatske narodne banke	52	3,561410111817	-0,911274616	+	n	n	n	n	N	A	A	N	
savjeta hrvatske narodne banke	51	3,403911330143	-3,968134876	+	n	n	n	n	N	A	A	N	
proračuna konsolidirane središnje države	50	2,790136643552	-2,71998006	+	n	n	n	n	N	A	A	N	
države članice europske unije	45	1,875181735881	1,912954023	+	n	n	n	n	N	N	A	N	
kamatne stope poslovnih banaka	30	1,430076898743	-1,605168532	+	n	n	n	n	A	N	A	N	
blagajničkih zapisa hrvatske narodne	29	1,833318940893	1,7194667833	+	n	n	n	n	A	N	A	A	
zapisa hrvatske narodne banke	29	-1,03658979205	-3,294482962	+	n	n	n	n	N	A	A	N	
državi članici europske unije	23	3,092927601727	1,8572072443	+	n	n	n	n	N	N	A	N	
tekućem računu platne bilance	21	2,943722825571	-1,570930038	+	n	n	n	n	A	N	A	N	
trezorskih zapisa ministarstva financija	21	2,043167859156	1,2715866353	+	n	n	n	n	A	N	N	N	
rast bruto dodane vrijednosti	18	-3,06258103096	-2,881042478	+	n	n	n	n	N	A	A	N	
kunski dio obvezne pričuve	18	1,521449593710	-2,09642756	+	n	n	n	n	A	N	A	N	
trezorske zapise ministarstva financija	18	2,385653692498	1,1478267613	+	n	n	n	n	A	N	N	N	
suglasnost hrvatske narodne banke	17	-0,99459403647	-2,503479917	+	n	n	n	n	N	A	A	N	
država članica europske unije	17	0,272084337167	1,4395271992	+	n	n	n	n	N	N	A	N	
obavljanje poslova platnog prometa	16	-0,37030543481	1,039455213	+	n	n	n	n	N	N	A	N	
ministarstva financija republike hrvatske	16	0,380667017353	1,5669235499	+	n	n	n	n	N	N	N	N	
inozemne aktive poslovnih banaka	14	0,77052614099	-0,844291399	+	n	n	n	n	A	N	A	N	
odobrenje hrvatske narodne banke	14	-3,21569599775	-2,781995134	+	n	n	n	n	N	A	A	N	
dijela devizne obvezne pričuve	13	-0,65770940303	-0,677574421	+	n	n	n	n	N	A	A	N	
međugodišnje stope rasta cijena	13	0,730021341733	0,6320884128	+	n	n	n	n	A	N	N	N	
nominalnoga efektivnog tečaja kune	13	1,454224182613	1,6905330030	+	n	n	n	n	A	A	N	N	
inozemni dug središnje države	12	0,281181752544	-0,229528736	+	n	n	n	n	A	A	A	N	
kamatnih stopa poslovnih banaka	12	-1,35133798054	-2,121011337	+	n	n	n	n	A	N	A	N	
guvernera hrvatske narodne banke	12	-0,72728293173	-0,545782403	+	n	n	n	n	N	A	A	N	
indeksa nominalnoga efektivnog tečaja	11	-1,26057785044	0,8295311956	+	n	n	n	n	N	A	A	N	
obavljanje funkcije člana uprave	11	-1,18966121950	-0,131162284	+	n	n	n	n	N	N	N	N	
investicijski portfelj vrijednosnih papira	11	1,302816326010	1,2090818414	+	n	n	n	n	A	N	A	N	
pričuve hrvatske narodne banke	10	-5,39888941663	-4,736565203	+	n	n	n	n	N	A	A	N	
prosječne aktivne kamatne stope	10	-1,15006039762	-2,422154398	+	n	n	n	n	A	A	A	N	
registra računa poslovnih subjekata	10	0,620021564079	-0,588732857	+	n	n	n	n	N	N	A	N	
valuta država članica emu-a	10	-0,51601514700	-1,936433800	+	n	n	n	n	N	N	N	Y	
rast bruto domaćeg proizvoda	9	-5,2655772108	0,2816672897	+	n	n	n	n	N	A	A	N	
rastu bruto dodane vrijednosti	9	-2,93659126198	-1,696416837	+	n	n	n	n	N	A	A	N	

Figure 1. Sample from top 138 tetragrams where frequency and linguistic filtering outperform MI

- (Boulis 2002) C. Boulis, 2002. *Clustering of Cepstrum Coefficients Using Pairwise Mutual Information*, Technical Report EE516, Electrical Engineering Dept. University of Washington, Seattle, 2002. (http://ssli.ee.washington.edu/ssli/people/boulis/ee516_report.pdf)
- (Church & Hanks 1990) K. W. Church & P. Hanks, 1990. *Word association norms, mutual information, and lexicography*, Computational Linguistics, 16, pp. 22-29.
- (Church et al. 1991) K. W. Church, W. Gale, P. Hanks, D. Hindle, 1991. *Using statistics in lexical analysis*, In U. Zernik (ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Lawrence Erlbaum Associates, 1991, pp. 115-164.
- (Daille 1994) B. Daille, 1994. *Study and Implementation of Combined Techniques for Automatic Extraction of Terminology*, In J. Klavans & P. Resnik (eds.), *The Balancing Act*, MIT Press, Cambridge MA, USA, 1994, pp. 49-66.
- (Daille 1995) B. Daille, 1995. *Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering*, In UCREL Technical Papers, Vol. 5, Dept. of Linguistics, Univ. of Lancaster, 1995.
- (Erjavec et al. 2003) T. Erjavec, C. Krstev, V. Petkevič, K. Simov, M. Tadić, D. Vitas, 2003. *The MULTEXT-East Morphosyntactic Specification for Slavic Languages*, In T. Erjavec, D. Vitas (eds.) *Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL2003*, Budapest, 2003, pp. 25-32.
- (Hatzivassiloglou 1994) V. Hatzivassiloglou, 1994. *Do We Need Linguistics When We Have Statistics? A Comparative Analysis of the Contributions of Linguistic Clues to a Statistical Word Grouping System*, In J. Klavans & P. Resnik (eds.), *The Balancing Act*, MIT Press, Cambridge MA, 1994, pp. 67-94.
- (Jacquemin 2001) C. Jacquemin, 2001. *Spotting and Discovering Terms through Natural Language Processing*, MIT Press, Cambridge, MA, USA, 2001.
- (Kilgariff & Tugwell 2002) A. Kilgariff & D. Tugwell, 2002. *Word Sketch: Extraction and Display of Significant Collocations for Lexicography*, In *Proceedings of the ACL-2001 Collocations Workshop*, Toulouse, France, pp. 32-38.
- (Loukachevitch & Dobrov 2003) N. V. Loukachevitch, B. V. Dobrov, 2003. *Knowledge-Based Text Categorization of Legislative Documents*, In *Proceedings of COMPLEX2003*, Budapest, Research Institute for Linguistics, Hungarian Academy of Sciences, 2003, pp. 57-66.
- (Manning & Schütze 1999) C. Manning, H. Schütze, 1999. *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge MA, 1999.

- (McEnery et al. 1997) T. McEnery, J.-M. Langé, M. Oakes, J. Veronis, 1997. *The Exploitation of Multilingual Annotated Corpora for Term Extraction*, In R. Garside, G. Leech, T. McEnery (eds.), *Corpus Annotation*, Longman, London, 1997, pp. 220-230.
- (Moguš et al. 1999) M. Moguš, M. Bratanić, M. Tadić, 1999. *Hrvatski čestotni rječnik*, Školska knjiga-Zavod za lingvističku Filozofskoga fakulteta Sveučilišta u Zagrebu, Zagreb, 1999.
- (Oakes 1998) M. Oakes, 1998. *Statistics for Corpus Linguistics*, Edinburgh University Press, Edinburgh, 1998.
- (Smadja 1993) F. Smadja, 1993. *Retrieving Collocations From Text: Xtract*, *Computational Linguistics*, 19(1), pp. 143-177.
- (Su et al. 1994) K.-Y. Su, M.-W. Wu, J.-S. Chang, 1994. *A Corpus-based Approach to Automatic Compound Extraction*. In *Proceedings, 32nd Annual Meeting of the ACL*, Las Cruces, NM, ACL, 1994, pp. 242-247.
- (Tadić 1994) M. Tadić, 1994. *Računalna obradba morfologije hrvatskoga književnoga jezika*, PhD Thesis, University of Zagreb, 1994.
- (Tadić 2002) M. Tadić, 2002. *Building the Croatian National Corpus*, In *Proceedings of LREC2002*, Grand Canaria, Spain, ELRA, 2002, pp. 441-446.
- (Tadić & Fulgosi 2003) M. Tadić & S. Fulgosi, 2003. *Building the Croatian Morphological Lexicon*, In T. Erjavec, D. Vitas (eds.) *Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL2003*, Budapest, 2003, pp. 41-46.
- (Tadić et al. 2003) M. Tadić, S. Fulgosi, K. Šojat, 2003. *The Applicability of Lemmatisation in Translation Equivalents Detection*, In *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, University of Birmingham Press, Birmingham, 2003, pp. 195-206.
- (Tiedemann 2000) J. Tiedemann, 2000. *Extracting Phrasal Terms Using Bitext*, In K.-S. Choi (ed.), *Proceedings of the Workshop on Terminology Resources and Computation, LREC2000*, Athens, Greece, ELRA, 2000, pp. 57-63.
- (Vintar 2000) Š. Vintar, 2000. *Using Parallel Corpora for Translation-oriented Term Extraction*, *Suvremena lingvistika* 49-50 (24/1-2), pp. 143-152.
- (Vintar 2002) Š. Vintar, 2002. *Avtomatsko luščenje izrazja iz slovensko-angleških vzporednih besedil*, In T. Erjavec & J. Gros, 2002, *Jezikovne tehnologije/Language Technologies*, SDJT-Jozef Stefan Institute, Ljubljana, 2002, pp. 78-85.