The 16th INTERNATIONAL DAAAM SYMPOSIUM
"Intelligent Manufacturing & Automation: Focus on Young Researchers and Scientists"
19-22nd October 2005

ATRIFICIAL INTELLIGENCE IN LEXICOGRAPHY: A CROATIAN ENCYCLOPAEDIC DICTIONARY EXAMPLE

Dembitz, Š.; Jojić, Lj. & Pavlek, J.

Abstract: *This paper gives a short overview of the possibility of applying the AI (Artificial Intelligence) methodology in lexicography. Tools and methods, originally developed for other NLP (Natural Language Processing) purposes, were modified in order to find what is missing at the entry side of the Croatian Encyclopaedic Dictionary. Finally, the paper is supported with a perception of a possible further development of the methodology presented here, all in order to fulfil the closure criteria considered as a major lexicographic demand on any dictionary.*

Key words: *artificial intelligence, closure criteria, lexicography, lexicology*

1. INTRODUCTION

The *Croatian Encyclopaedic Dictionary* (from now on referred to as HER, 1st edition 2002, 2nd edition 2004-05) is a result of a team work that developed further its two main sources: the *Dictionary of Croatian Literary Language* (Anić, 1998) and the *Dictionary of Foreign Words* (Anić & Goldstein, 2000). Its structure differs significantly from any dictionary ever published in Croatia. The words are being grouped in clusters. For the first time the onomastic is being introduced as a part of a dictionary as well as the etymology for originally Croatian words. The public reception of the work was very positive. The critics welcomed it, but not without critical remarks. The criticism provoked our interest for non-conventional ways of improvement. Convinced that a dictionary like HER is a crucial cultural fact for every nation, we decided to apply the AI (Artificial Intelligence) technology on it, in order to improve what can be improved in a short time and with limited man-power resources.

The paper is organized in five sections. Section 2 defines the closure criteria in lexicography. Section 3 describes the methodology we developed in order to fulfil closure criteria, accomplished with some metric results of its application on HER. Finally, Section 4 points out what should be done in the future in order to satisfy all what is meant under closure criteria.

2. CLOSURE CRITERIA

Simply stated, closure criteria for a dictionary mean that everything what occurs on the right side of a dictionary must be listed on the left, entry side of the same dictionary. It is a golden rule of a good lexicography. Being the golden rule closure criteria are difficult to be fulfilled.

There are very few dictionaries of Croatian language. Their authors, though prominent linguists with experience in working on dictionaries, usually lack the knowledge of the lexicography proper. As a result they produced dictionaries without the right lexicographical solutions. These are hardly to be found in contemporary Croatian lexicology or grammars. Some recent works in the field of lexicology do not find their way to the authors of the dictionaries (Tafra, 2003a, 2003b). The new dictionaries are being compiled from the older ones, so the old mistakes perpetuate themselves. No dictionary ever produced in Croatia can claim to satisfy closure criteria.

A dictionary, when published, is an object of public criticism. However, can a dictionary become a subject of criticism, or better formulated, an object of self-criticism? We are convinced it can, if their editors apply in a proper way the computer technology.

Starting from closure criteria we applied on HER some program tools initially developed for *Hascheck* (Dembitz et al., 1998, 1999), in order to see what is in definitions but is not listed as a dictionary entry. Since the Croatian language is highly inflected language it was not a trivial task. The next section describes what we have achieved and gives a hint how a dictionary can be used for self-improvement.

3. DICTIONARY - SOURCE FOR SELF-IMPROVEMENT

HER, without its onomastic region, is a corpus of 1.7 million tokens. It has something more then 120,000 dictionary word entries given in lemmatized form (verbs in infinitive, nouns in nominative, etc.). It means that the right side, the definitions, contains 1.6 million tokens with a certain distribution of word-types. We found reasonable to look first for the word-types in definition zone having the frequency of occurrence equals 1. These are the most interesting word-candidates for improving the left, entry side of the dictionary. The definitions contain 97,187 unique word-types (UWT) belonging to the class of so-called common words (common word is a word that can be written in minuscules, with starting majuscule or in all majuscules without changing the meaning).

Almost all of those 97,187 word-types were not lemmatized but are the inflected forms of Croatian words. In order to become able to see what worth in this set is really for a lexicologist, *Hascheck's* tagging algorithm (Dembitz et al., 2003) was adapted to cope with flections having a set of lemmatized words as the base for connecting them. We applied a strict engineering approach: minimum of programming for the maximum of output. The result of this approach is given in Table 1.

First column in the Tab. 1 gives the global distribution of the definition word-types having frequency equals 1 (UWT_1). Second column gives the distribution of those word-types from the column 1 not connected by the tagger with entry words (UWT_2). Third column is the compression rate:

$$CR = \left(1 - \frac{UWT_2}{UWT_1}\right) \cdot 100\,[\%] \qquad (1)$$

The average CR is nearly 85%; it is estimated as a good compression. CR varies from letter to letter. Therefore the letter rows in Tab. 1, where the corresponding CR is under 80%, are bolded. These rows are **i, n, o, p, u** and **z**. Every Croatian lexicographer would tell these letters are the most problematic when preparing a dictionary. It is so because many words start with short prefixes *iz-, na-, ne-, o-, od-, po-, pod-, pre-, pri-, pro-, u-* and *za-*. Here the program confirms the lexicographer's experience. Simply speaking, the main contribution for improving HER entries in terms of closure criteria is expected from these six letters.

	Unique word-type (UWT $_1$) distribution		Non-connected (UWT $_2$) distrib.		C-rate (CR)
A	4 831	5.0%	337	2.3%	93.0%
B	4 505	4.6%	332	2.2%	92.6%
C	1 195	1.2%	140	0.9%	88.3%
Č	926	0.9%	116	0.8%	87.5%
Ć	131	0.1%	13	0.1%	90.1%
D	4 507	4.6%	626	4.2%	86.1%
Dž	100	0.1%	3	0.0%	97.0%
Đ	93	0.1%	4	0.0%	95.7%
E	2 355	2.4%	157	1.1%	93.3%
F	1 948	2.0%	159	1.1%	91.8%
G	2 870	3.0%	236	1.6%	91.8%
H	2 297	2.4%	143	1.0%	93.8%
I	**4 022**	**4.1%**	**868**	**5.9%**	**78.4%**
J	1 484	1.5%	108	0.7%	92.8%
K	7 386	7.6%	663	4.5%	91.8%
L	2 045	2.1%	158	1.1%	92.3%
Lj	236	0.3%	30	0.2%	87.3%
M	4 637	4.8%	438	3.0%	90.6%
N	**5 537**	**5.7%**	**1 457**	**9.8%**	**73.7%**
Nj	74	0.1%	8	0.1%	89.2%
O	**5 145**	**5.3%**	**1 337**	**9.0%**	**74.0%**
P	**13 046**	**13.4%**	**2 844**	**19.2%**	**78.2%**
R	4 353	4.5%	730	4.9%	83.2%
S	7 881	8.1%	1 199	8.1%	84.8%
Š	1 513	1.6%	207	1.4%	86.3%
T	3 804	3.9%	391	2.6%	89.7%
U	**3 249**	**3.3%**	**912**	**6.2%**	**71.9%**
V	2 531	2.6%	287	1.9%	88.7%
Z	**3 801**	**3.9%**	**842**	**5.7%**	**77.8%**
Ž	685	0.7%	81	0.5%	88.2%
Total:	97 187		14 826		84.7%

Table 1. Distribution of non-connectable definition word-types

To estimate the usefulness of the compression we have done, a lexicological evaluation was performed on samples represented in row 1 (words starting with the letter *a-*) and row 30 (words starting wit the letter *ž-*), respectively. Both samples are relatively small (337 and 81 word-types left non-connected after compression, respectively) and with a high CR (93% and 88%, respectively). Here are the results of the lexicological evaluation:

- among 337 word-types starting with *a-*, not connected with dictionary entry words, 134 (40%) are declared to become new entry words;
- among 81 word-types starting with *ž-,* not connected with dictionary entry words, 9 (11%) are declared to become new entry words.

The rest of non-connected word-types are the so called indirect entries, described in grammar zone of HER.

The lexicological analysis of these two samples demonstrated that 34.2% of non-connected unique word-types from the definition zone of HER are usable for improving the left side of the dictionary in terms of closure criteria. Extended on the whole sample from Tab. 1 it means at least 5-6 thousand new entries for HER, or an enlargement of 5% in the dictionary entry numbers. A lexicographer using only conventional methods knows what it means in terms of effort and working hours. Our approach drastically reduces this effort (equals money) by using the dictionary corpus and appropriate AI methods.

Having done it for the word-types with frequency equals 1, we extended our methodology on the word-types with higher frequencies in definition zone. We found 6.999 such word-types not connectable with entry zone. The most frequent one is *sl.* (abbreviation for *similar* and *Slovakian* in HER), with the frequency equals 4.948. Our «mistake» was that we didn't include the table of lexicographical abbreviations in our tagging process. However, from the lexicologist point of view the most interesting finding were the verbs in infinitive (extractible automatically), used in definitions but not in entry zone. Exactly 129 such verbs were found.

4. FURTHER RESEARCH AND DEVELOPMENT

The methodology we developed so far for the self-improvement of HER, in order to satisfy the closure criteria, is based on the treatment of isolated words. To become able to seize the differences in word meaning regarding their usage, and so to improve the techniques we have developed up to now, our research is directed towards multiple alignment, i.e. treating words in context and comparing their contextual usage metrically. We are just on the very beginning of this research. In this moment we cannot say how far the research will come, and what will be the practical output of it. However, we think it's the only way to close our methodology for fulfilling closure criteria in the lexicography.

5. CONCLUSION

We are well aware of the fact that, in the way of methodology, we made a progress in Croatian lexicology as well as in lexicography. This progress was enabled by introducing AI methods adequately into the field. Some results are presented here, and they are very encouraging. However, we are not fully satisfied with what we have done. Our achievement is a result of applying the existing AI tools, slightly modified, in a new manner and for a new purpose. A real breakthrough will be done when one can process the entire lexicon context in the AI manner. Therefore, we now focus our research and development on multiple alignment methods with the scope of applying them in lexicology and lexicography, primarily having HER in mind. Its cultural relevance is too high to even think of giving up in any moment.

6. REFERENCES

Anić, V. (1998). *Rječnik hrvatskoga književnoga jezika (Dictionary of Croatian Literary Language, 3th edition)*, Novi Liber, Zagreb, Croatia.

Anić, V., Goldstein. I. (2000). *Rječnik stranih riječi (Dictionary of Foreign Words, 2nd edition)*, Novi Liber, Zagreb, Croatia.

Dembitz, Š., Knežević, P., Sokele, M. (1998). Learning Words – A Cognoelectrical Analogy, *Proc. 9th Artificial Intelligence / Cognitive Science Conference – AISC'98*, pp. 47-54, Dublin, Ireland, August 19-22, 1998.

Dembitz, Š., Knežević, P., Sokele, M. (1999). Hascheck – The Croatian Academic Spelling Checker, *Applications and Innovations in Expert Systems VI* (R. Milne, A. Macintosh & M. Bramer, Eds.), pp. 184-197, Springer: London-Berlin-Heidelberg-New York-Barcelona-Hong Kong-Milan-Paris-Santa Clara-Singapore-Tokyo.

Dembitz, Š., Knežević, P., Sokele, M. (2003). Knowledge Acquisition and Energy: A Case Study, *Proc. IASTED International Conference: Artificial Intelligence and Applications*, pp. 309-314, Innsbruck, Austria, February 10-13, 2003.

Tafra, B. (2003a). Preispitivanje Hrvatske jezične norme (Auditing of the Croatian Language Norm), *Jezik*, 50(2):48-58.

Tafra, B. (2003b). Koji i kakvi rječnici (Which and What Dictionaries, manuscript for oral presentation), presented on *Hrvatski jezični krug*, April 15, 2003.