Evaluation of Open-Source Online Dictionaries

Dina Crnec and Sanja Seljan
Department for Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
Phone: (385) 1-600 2350 Fax: (385) 1-600 2431 E-mail: dcrnec@ffzg.hr, sseljan@ffzg.hr

Abstract – The paper stresses the importance of promoting multilingualism in today's society and the symbiosis of technology and education in general. The emphasis is put onto necessity for development of online dictionaries of open-source type as a part of language technologies, seen through the aspect of tools for translation and language learning. The objectives of the study are as follows: comparative analysis and evaluation of Croatian-English and English-X dictionaries according to specific quality rank, series of questions related to examinees' previous education – place and preferred ways of education, and finally, examinees' self-evaluation of four language skills according to the *European language levels Self Assessment Grid.* The research was conducted among the students of information sciences at the Faculty of Humanities and Social Sciences, University of Zagreb, using the method of survey. The results have been analyzed and further recommendations and guidelines given.

I. INTRODUCTION

The paper starts from the fact that today's scientific and cultural cooperation, international multilingual communication and recent changes in education require the creation of online electronic language resources, i.e. different types of dictionaries.

The information society of today pushes upon following trends and constant technology enhancements, thus launching the so-called *e-trend* in many spheres of life, e.g. e-learning or e-business. This also implies considerable changes in the domain of media, communication, information resources, educational and learning methods etc. where Internet has a leading role. Also, it is important to mention the usage of language technologies containing i.e. multimedia online dictionaries in teaching and learning processes.

Nowadays, multilingualism is no longer just an additional skill but a need, thus language technologies are emerging as a constantly developing interdisciplinary field. In the paper, the focus is on language resources, namely online dictionaries of open-source type, seen from the users' perspective regarding translation and language learning.

II. CLASSIFICATION OF DICTIONARIES

"The electronic dictionary is not a mere electronic version of a book-form dictionary (i.e. machine-readable dictionary, MRD). It is a dictionary for computer processing of natural language, e.g. word processing and machine translation." (Electronic Dictionary Research, 1994). Electronic dictionaries could be seen as the biggest component of Machine Translation (MT) systems considering the amount of information they provide, although the ones included into MT systems require less detailed information, such as pronunciation. The main differences between electronic and traditional book-form dictionaries, besides stating the obvious, certainly lie in format, coverage, term precision and linguistic descriptions. Electronic dictionaries require the same quality and information details as that of book-form dictionaries, but that is very often not the case.

When speaking of electronic dictionaries' classification, there are a few classifications that will be presented further on. Firstly, we could roughly divide them into three categories, according to the number of languages they include: monolingual, bilingual and multilingual electronic dictionaries.

Monolingual electronic dictionaries are of best quality among the other two categories, and very much alike the book-form dictionaries. They are mostly compiled and organized by editors. Very often, monolingual electronic dictionaries can be found as an integral part of bigger sites, encyclopedias or other dictionaries. They can have integrated spelling checkers and screen similar terms if a query could not be found, with the results showing up as links to other terms and definitions. Besides online dictionaries, the ones of the best quality are certainly CD-ROM dictionaries. These are made as similar as possible to book-form dictionaries, with multimedia add-ons such as interactive exercises, thesauri or collocation information.

Secondly, there are bilingual and multilingual electronic dictionaries which differ very much from the book-form dictionaries and are of considerably lower quality than monolingual dictionaries. A good feature that can often be found is the sound/pronunciation option, together with occasional spelling checkers, but very often there are problems with the keyboard and the recognition of alternative characters.

Multilingual electronic dictionaries basically resemble to the bilingual type of electronic dictionaries by quality, but usually contain a very narrow selection of terms per language, depending on the number of languages included. Here, the concept of free creation appeared first, also present within other types of web resources. Users can contribute to the system by correcting obvious errors, giving comments and adding up content to the dictionary, which in the end modifies the system itself. Tools and resources in general change every day, but at the same time they ask for much more caution, evaluation and revision as the quality level is generally on a quite low level.

Another way of classifying electronic dictionaries would be according to the type of media: CD-ROMs and online dictionaries, both either open-source or requiring registration and/or payment. CD-ROMs usually contain a lot of information, including sound files for pronunciation,

language learning materials (for English mostly) and even images. Their most frequent downside is not so user-friendly interface design as well as pricing. On the other hand, online dictionaries are free by nature, which also implies some distracting advertisements. Users are mostly looking for idiomatic expressions, where online dictionaries enable them to check the context of desired terms and examples.

Yet another way of classifying electronic dictionaries would be into general and specialized dictionaries, depending on the topic and coverage. General dictionaries provide information about the most common words of a language. However, new words are often not being included in these dictionaries until they become very common, so those words should be sought for in a dictionary of slang or in another specialized dictionary. Specialized dictionaries contain words useful for specific groups of people, e.g. learner's dictionaries for learning a language, technical, medical or computer dictionaries, thesauri, dictionaries of idioms/collocations, dictionaries of slang, visual dictionaries, dictionaries of synonyms and antonyms etc.

Whether a dictionary is printed or electronic, online or offline, it is of no importance for the content of the dictionary but merely a form of presentation. Dictionaries can contribute to better understanding of language, asking for good organization, excellent coverage in the domain, consistency and/or multimedia elements. Considering all of the aforementioned types of dictionaries and given the fact that open-source online dictionaries are the most frequently used type of dictionaries today, especially among younger population, we limited our research to open-source online dictionaries only.

III. SITUATION IN CROATIA

There have been significant changes in educational and organizational spheres of life in Croatia. Multilingualism and multiculturalism are slowly becoming a piece of everyday life, especially when referring to education.

At the time being, some faculties in Croatia are introducing new final changes in curricula and module organization in order to level our programs with the ones in the European Union (EU). Changes in our educational system also include promoting international student and staff mobility, usage of e-learning systems, introduction of new ways of teaching and learning etc. At the Faculty of Humanities and Social Sciences – University of Zagreb, there are currently 23 language studies, among 75 possible study groups.

When speaking of electronic resources for Croatian language, there are, although not many and of doubtful quality, certain resources both online and offline. Today, the most developed language technologies are for English language respectively, which also points out to the significant gap between the existing technologies for English and other world languages, not to mention smaller, generally under-resourced languages. For Croatian language specifically, there are some available resources such as corpora, different types of dictionaries, translation memories, terminology databases, thesauri and lexicons in progress. Web data concerning Croatian language is generally quite scattered on different addresses on the WWW. The existing open-source online dictionaries for Croatian language mean, in most cases, translation Croatian-English and vice versa. These mostly use the same database (by phD Goran Igaly) resulting in a very similar quality of translated words (e.g. *EH.Web*, *Bosiljak*, *LangTran*). Other open-source online dictionaries to be mentioned are *e-Rječnik* and *Hrvatski rječnik* (both multilingual dictionaries), as well as *EU Dict* (a collection of online dictionaries for different languages) and the *Babylon* collection of dictionaries for Croatian language, which has a trial and a payable version. There are certain bilingual dictionaries, such as *croDict* (Croatian-German) or *Kadingira* (Croatian-Slovene, also including dictionaries for a few other languages), a monolingual Croatian dictionary *Hrvatski jezični portal*, some specialized dictionaries (Internet dictionary, dictionary of communications technology etc.), *Webster*'s Croatian-English dictionary with multilingual thesaurus translation, a multilingual terminology database including Croatian language *Evroterm* etc.

Other than that, there is a project called "Croatian Dictionary Heritage and Dictionary Knowledge Representation" led by ph.D Damir Boras, resulting in a web portal called *Croatian Old Dictionary Portal*. It contains many digitized dictionaries, several of which are available for searching through and browsing, as well as cross-searching. Also, there is an institutionally supported portal for the Croatian language called *Language Technologies for Croatian Language (JTHJ)*, endorsed by the Institute for Linguistics at the Faculty of Humanities and Social Sciences, University of Zagreb. Its most interesting part corresponding to our interests is certainly the list of dictionaries for Croatian in combination with many languages, out of which we already mentioned some. The portal itself could use a more often content refresh though, since it is a web location containing a very large number of resources for Croatian language, not only dictionaries but also corpora, language tools, speech processing, information about projects, organizations and conferences etc.

IV. RESEARCH: RELATED WORK

Having studied related work in the area of evaluation of electronic dictionaries, we have noticed the lack of freely publicly available information on such researches and their results. One of the first projects related to dictionary analysis was *EDR - electronic dictionary project* (1986-1995) in Japan. The project plan was drawn up for the development of dictionaries for usage in Natural Language Processing (NLP). There is one very important guideline given during the project back in 1994, that is still not being completely followed and should be when referring to constructing a large-scale electronic dictionary – paying special attention to providing a very large-scale corpus to extract linguistic information and verify data described in the electronic dictionary.

Another research of this kind was an overview and comparative analysis of trends in electronic dictionary research and development in Europe, where lexical database systems in Europe have been studied and compared among them and to *EDR*. Also, the existing European lexical workstations at the time, monolingual and multilingual, have been studied (Sérasset, 1994).

Besides these, there was a more recent research on dictionary usage in the context of foreign language learning, published in the form of a book containing reports on a series of studies which the author had conducted in the past fifteen years (Tono, 2001). It brings together some of the findings of studies on dictionary users and shows how research into dictionary usage can contribute to the improvement of dictionary design and the clarification of issues in language learning.

There was a recent research regarding evaluation of online dictionaries for language learning (Kawasaki, Miyaji, Yamaguchi), conducted in 2008. The authors examined educational effectiveness and efficiency of online dictionaries by comparing two types – a frame style and a pop-up style. The results of their study on English-Japanese translation showed that the readability was more accurate with the frame style dictionary than with a pop-up dictionary, finally resulting in a conclusion that frame-type online dictionaries would become a very powerful learning tool in e-Language Learning (eLL) to enhance learners' reading ability.

Furthermore, according to a study on evaluation of electronic translation tools through quality parameters, conducted in 2009 (Kučiš, Seljan, Klasnić), the introduction of additional Computer-Assisted Translation (CAT) tools at the statistically significant level influences the quality and the consistency of translation.

Lastly, there was a recent case study on improvements of dictionaries and suggestions by *Evroterm* (Željko, 2009). The author presents some possibilities for improving electronic dictionaries from a translator's point of view and suggests continuous improvements in the content and technical features instead of constantly new dictionary projects. Further on, the paper presents some possibilities for designing electronic dictionaries in order to overcome book limitations; from full-text and fuzzy search, usage of corpora for retrieving examples of usage and terminological analysis of the text to be translated, to continuous improvements of software and data, as well as dictionary copyright protection with digital certificates.

V. RESEARCH: SURVEY

Having examined previous studies and in order to come up with concrete results on the quality of open-source online Croatian and English language dictionaries, as well as the information on language skills, we conducted a research using the method of survey among the students of information sciences (major and minor study) at the Faculty of Humanities and Social Sciences – University of Zagreb.

The survey was conducted online, using the *Omega* e-learning system, a Croatian version of *Moodle* (Modular Object-Oriented Dynamic Learning Environment). The students-examinees were asked to download the survey to a physical location, fill it out and upload back through *Omega*. The primary goal was to gather information on examinees' perspective of different tools for translation and language learning, and apply it to our hypothesis stated at the beginning of this paper – the necessity for development of online dictionaries of open-source type as a part of language technologies.

In this research, the sample for the survey consisted of 74 students, according to stratified quote convenience sample, belonging to:

- Two courses of all together 28 students who have already attended courses related to the usage of language resources, i.e. online dictionaries: one course on the 3rd year of undergraduate study (18 students) and another on the 1st year of graduate study (10 students),
- One course on the 1st year of undergraduate study (35 students) who still haven't had a course related to language technologies,
- One random sample consisting of students on the 3rd year of undergraduate study and the 1st year of graduate study (11 students) in which some of them had courses related to the usage of dictionaries for translation purposes or language learning.

The examinees were given a survey of nine questions, divided into several categories:

- information on previous language learning (place of language acquisition, type of preferred language learning methods, reasons for language learning, preferred open-source online dictionaries),
- self-evaluation of language skills according to the *European language levels Self Assessment Grid*, a common European framework of reference for languages covering the areas of learning, teaching and assessment,
- dictionary evaluation relating to the evaluation of Croatian-English and English-X dictionaries, according to the following criteria: translation adequacy, dictionary usability, number of meanings per word, order of meanings' importance, information per word (synonyms, antonyms, examples), sound record and quality of explanations per meaning.

The examinees compared and evaluated three to four dictionaries for Croatian-English and also three to four for English-X language. The evaluation scale used in this evaluation process was from 1 (insufficient) to 5 (excellent).

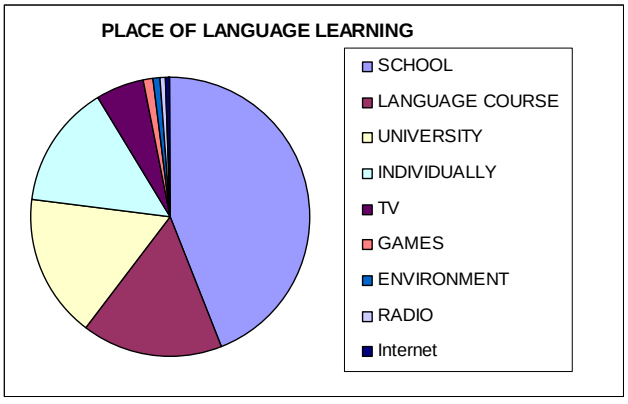

Fig. 1. Place of language learning.

Regarding the purpose of language learning, 62 examinees answered related to work, 23 related to a translation job, 25 related to studies and 17 answered other.

Among 74 students, 42 of them prefer an individual type of learning, 38 prefer group learning, 24 classic ways of learning, 4 e-learning, 3 combination learning etc.

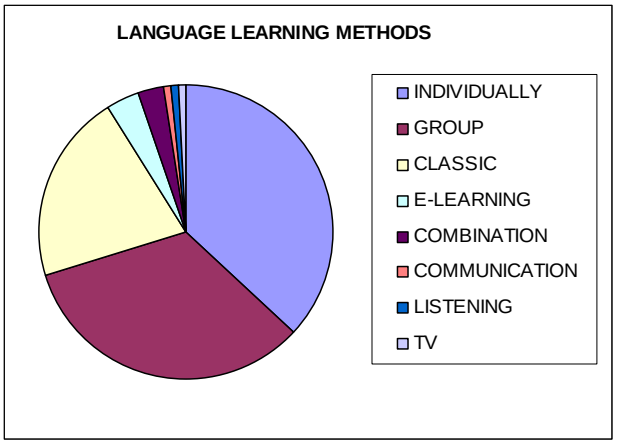

Fig. 2. Language learning methods.

The next section in the survey is related to self-evaluation of four language skills: reading, writing, listening and speaking. 74 students indicated 182 languages all together, i.e. the skills for 18 different languages: English (72); German (44); Italian (16); Spanish (12); French (11); Russian (6); Czech, Slovak, Slovenian and Japanese (3); Dutch (2) etc. According to the data, the students of information sciences use at a certain level 2.459 languages per person, as showed in Table 1.

TABLE I
EUROPEAN LANGUAGE LEVELS SELF ASSESSMENT GRID

	LISTENING			READING			SPEAKING			WRITING		
European language level	C2/C1	B2/B1	A2/A1	C2/C1	B2/B1	A2/A1	C2/C1	B2/B1	A2/A1	C2/C1	B2/B1	A2/A1
English	**49**	**17**	1	**48**	**18**	1	**37**	**29**	1	**38**	**28**	1
German	**11**	**12**	**22**	11	**14**	**19**	8	11	**25**	9	9	**27**
Italian	3	4	10	3	3	11	3	2	12	3	2	**12**
Spanish	1	7	4	1	7	4	1	3	8	1	3	8
French	0	4	8	0	4	8	0	2	10	0	3	9
Russian	4	1	1	3	2	1	3	2	1	3	2	1
	68	45	46	66	48	44	52	49	57	54	47	58

According to the data gathered in this study, it is obvious to conclude that English language is the most used language. Listening and reading skills at the level of proficient user (C2, C1) count for 49 students, while speaking and writing skills are mostly represented at the level of proficient user and independent user (B2, B1).

The second most used language is German, which is more equally self-evaluated among levels and skills. Levels of proficient user and independent user are closely represented for all skills, but the writing skill is mostly represented by the basic user's level (A2, A1).

Italian language is the third most used language, where basic user's level is mostly represented. Spanish language is more frequently self-evaluated at B levels for listening and reading skills, and for speaking and writing skills at A levels. For French language there aren't any advanced proficient users for any of the skills; for all skills, French is present at A levels. Russian language is more divided between A and B levels, except for listening skills which are mostly self-evaluated at C levels.

Regarding the question related to the type of online dictionaries used, it indicates that 49 students use monolingual online dictionaries, 38 bilingual dictionaries and 24 of them multilingual dictionaries.

One of the examinees' tasks was to evaluate three or four suggested Croatian-English online dictionaries:

http://www.rjecnik.net/,
http://www.eudict.com/,
http://www.e-rjecnik.net/,
http://www.taktikanova.hr/eH/eh.asp.

They were also asked to evaluate three or four suggested English-X language dictionaries:

http://www.collinslanguage.com/,
http://www.wordreference.com/,
http://dictionary.reverso.net/,
http://dict.tu-chemnitz.de/.

Apart from these suggested dictionaries, they could choose any other dictionary according to their preferences. Dictionaries have been evaluated according to the following criteria through average grades:

- translation adequacy (TA),
- dictionary usability (DU),
- number of meanings per word (NMW),
- order of meanings' importance (OMI),
- number of information per word – synonyms, antonyms, examples (NIW),
- sound record (SR),
- quality of explanations per meaning (QE).

TABLE II
AVERAGE GRADES PER CRITERIA

CRITERIA	TA	DU	NMW	OMI	NIW	SR	QE
AVERAGE - ENGLISH	4,04	3,96	3,85	3,76	3,71	2,27	3,73
AVERAGE - CROATIAN	3,83	3,92	3,68	3,38	2,60	1,28	2,59

The correlation between these two groups of evaluated dictionaries is 0.89978.

Values for English-X dictionaries are higher for every criteria than for Croatian-English dictionaries and range from 3.70860 to 4.04034, except for the sound record with

an average grade of 2.26869. Higher grades for Croatian-English dictionaries are for the criteria on translation adequacy (TA), dictionary usability (DU) and for number of meanings per word (NMW), ranging from 3.68291 to 3.83460, probably due to the necessity for the usage of online dictionaries. The criteria on number of information per word – synonyms, antonyms, examples (NIW) and quality of explanations per meaning (QE) ranges from 2.58517 to 2.60017. The lowest grade assigned to sound record (SR) is 1.28123.

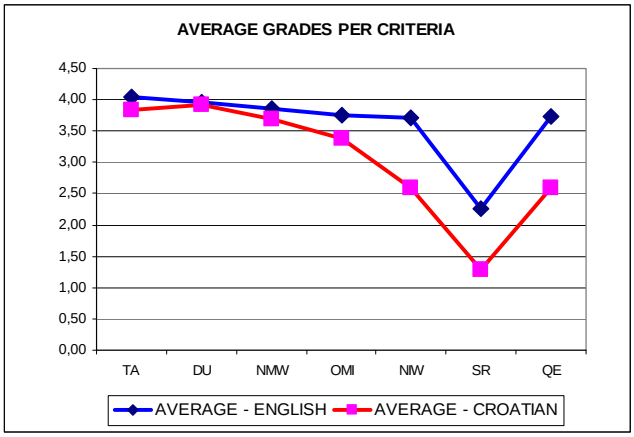

Fig. 3. Average grades per criteria.

As seen from the data, there is an obvious need for the development of online language resources. The students of information sciences speak in average 2.459 languages and frequently use open-source online monolingual and multilingual dictionaries. Croatian-English dictionaries are marked in every criteria with lower grades than English-X dictionaries, pointing to the need for good quality bilingual Croatian dictionaries.

VI. CONCLUSION

One of the main goals of this paper was to emphasize the necessity for development of online dictionaries of open-source type as a part of language technologies, seen through the aspect of tools for translation and language learning. Three different classifications of dictionaries were presented, according to the number of languages they include, types of media and topics/coverage. There was an overview of the current situation in Croatia regarding educational changes and electronic resources for Croatian language, including various existing open-source online dictionaries.

Related work has been presented in the form of certain projects and studies on the evaluation of electronic dictionaries. In order to come up with concrete results on the quality of open-source online Croatian and English language dictionaries, as well as the information on language skills, the authors conducted a research using the method of online survey among the students of information sciences.

As seen from the data gathered in this study, English language is the most used language and was marked with mostly proficient C and independent B levels. Our students indicated all together 182 languages, which makes approximately 2.5 languages per person. They very frequently use open-source online monolingual and multilingual dictionaries for different purposes. Values for English-X dictionaries were higher than for Croatian-English dictionaries, most probably due to the necessity for the usage of online dictionaries.

These findings show a strong need for the development of online language resources, especially the development of good quality bilingual Croatian dictionaries, in order to enable better understanding of the language and contribute to decreasing the gap between resourced languages (i.e. English) and under-resourced ones.

REFERENCES

[1] Boras, Damir. Croatian Old Dictionary Portal *http://crodip.ffzg.hr/default_en.aspx*

[2] Electronic Dictionary Research (EDR) Project. http://www.wtec.org/loyola/kb/c5_s2.htm

[3] European Commission Joint Research Centre. *http://langtech.jrc.it/*

[4] European Commission Multilingualism. *http://ec.europa.eu/education/languages/eu-language-policy/index_en.htm*

[5] Fuentes Morán, T; García Palacios, J; Torres del Rey, J. Algunos apuntes sobre la evaluación de diccionarios. REV - RL - Vol. 11, (2004-2005) pp. 69-80. (in Spanish)

[6] Ide, N; Véronis, J. Knowledge extraction from machine-readable dictionaries: An evaluation. Springer, 2006.

[7] Ide, N; Véronis, J. Text, Speech and Language Technology. Springer, 2010.

[8] Jezične tehnologije za hrvatski jezik. *http://www.hnk.ffzg.hr/jthj/* (in Croatian)

[9] Kawasaki, Y; Miyaji, I; Yamaguchi, H. & Y. Evaluation of On-line Dictionaries for Language Learning. *Proceedings of Society for Information Technology & Teacher Education International Conference 2008* (pp. 5242-5247).

[10] Kučiš, V; Seljan, S; Klasnić, K. Evaluation of Electronic Translation Tools Through Quality Parameters. // INFuture 2009 – Digital Resources and Knowledge Sharing. Zagreb. 2009. pp. 341-351.

[11] Sérasset, Gilles. Recent Trends of Electronic Dictionary Research and Development in Europe. Taylor & Francis US, 1998.

[12] Sustav učenja na daljinu Omega. *http://omega.ffzg.hr/* (in Croatian)

[13] Tono, Yukio. Research on Dictionary Use in the Context of Foreign Language Learning. De Gruyter Mouton, 2001.

[14] Types of Dictionary – Advanced Dictionary Skills Program. *http://elc.polyu.edu.hk/advdicts/types.htm*

[15] Željko, Miran. Improvements of Dictionaries – Suggestions by Evroterm. // INFuture 2009 – Digital Resources and Knowledge Sharing. Zagreb. 2009. pp. 269-278.