How to Evaluate and Raise the Quality in a Collaborative Lexicographic Approach

Dan Cristea[1,2], Corina Forăscu[1,3], Marius Răschip[1], Michael Zock[4]

1 "Alexandru Ioan Cuza" University of Iaşi
2 Romanian Academy, Iaşi
3 Romanian Academy, Bucharest
4 Université de Marseille à Luminy
E-mail: {dcristea,mraschip,corinfor}@info.uaic.ro, Michael.Zock@lif.univ-mrs.fr

Abstract

This paper focuses on different aspects of collaborative work used to create the electronic version of a dictionary in paper format, edited and printed by the Romanian Academy during the last century. In order to ensure accuracy in a reasonable amount of time, collaborative proofreading of the scanned material, through an on-line interface has been initiated. The paper details the activities and the heuristics used to maximize accuracy, and to evaluate the work of anonymous contributors with diverse backgrounds. Observing the behaviour of the enterprise for a period of 6 months allows estimating the feasibility of the approach till the end of the project.

1. Introduction

The Web offers nowadays many digital dictionaries[1]. We could even say that the dictionaries are going Web. This change in support has also caused changes concerning the way of contributing content. Since humans speak languages, why not ask them to fill the slots of the dictionary's microstructure? Accepting the idea that ordinary people could contribute in building a dictionary on the basis of their knowledge of a certain language implies a dramatic change of view concerning the way dictionary content is provided. Terms like "crowdsourcing"[2] and "digital sharecropping" (Zimmer, 2007) have been coined and start to be used also in linguistics. Recently, there has been a raise of interest in acquiring lexicographic data by using free, large scale, work, over the Web[3]. The reasons for this interest are probably economic: productive costs need to be reduced. The method is not uncontroversial, because a collaborative approach might be dangerous (almost malpractice) as incompetent contributor may produce a lot of noise, hard to get rid of. We will discuss here how to assure quality while thousands of unknown people provide data, how to motivate them, and how to evaluate their work.

The practical setting of our approach is the conversion of the paper version of a very large dictionary into its electronic format: the Thesaurus Dictionary of the Romanian Language, an explanatory dictionary built under the auspices of the Romanian Academy since 1913 (once finished, in 2008, it will include 33 volumes, more than 15,000 pages, about 175,000 entries and more than 1,300,000 examples). The dictionary was created in the traditional pencil-and-paper way. It includes an index for more than 2,500 volumes of the Romanian literature. eDTLR is the name of the digital form of the Dictionary. It includes the sources in digital form and the software to access them.

This paper is organized as follows. Section 2 presents the two parts of the Thesaurus Dictionary of the Romanian Language, as well as the eDTLR project's objectives. Section 3 presents thee important issues in the collaborative approach of building eDTLR: how to involve many contributors, how to obtain accuracy, and how to evaluate their work. Section 4 presents the current state in the development of eDTLR, first results and, based on them, prefigures the near future, and the last section gives conclusions.

2. DA, DLR and eDTLR

Building the content of the Thesaurus Dictionary of the Romanian Language took almost one century. The old series, known as the Dictionary of the Academy (DA) included 5 volumes with 3,146 pages and 44.890 entries, and has been developed between 1913 and 1947 by the Romanian Academy. After an interruption, the work was restarted in the middle of the 7th decade of the last century with the new series, known as the Dictionary of Romanian Language (DLR). Is it expected that the dictionary will be finalised at the end of 2007. In all, DA and DLR will have 33 volumes, more than 15,000 pages, about 175,000 entries and more than 1,300,000 examples. The dictionary

---

[1] There are 800 dictionaries in 160 languages at http://ling.kgw.tu-berlin.de/call/webofdic/diction4.html. See also Digital dictionaries of South Asia or ARTFL at www.lib.uchicago.edu/efts/ARTFL/projects/dicos/
[2] http://en.wikipedia.org/wiki/Crowdsourcing
[3] Examples are:
- Wiktionary (www.wiktionary.org) – a multilingual collection of free dictionar-ies in over 150 languages;
- the Kamusi project, (www.kamusiproject.org) – a Swahili-English dictionary;
- the Papillon project, (www.papillon-dictionary.org/Home.po) – a multilingual dictionary for ori-ental and western languages;
- the Inuktitut Living Dictionary (www.livingdictionary.com/backgroundandhistory.jsp);
- the Online Slang Dictionary (http://onlineslangdictionary.com)

was created in the traditional pencil-and-paper way, including the index on more than 2,500 volumes of the written Romanian literature, till the nineties, when for editing and publication the lexicographers started to make use of computers.

eDTLR ("e" stands for electronic versions) represents the name of the digital form of DA+DLR (1965), including its sources in digital form and the software to access them, as well as the name of a three years project. The project focuses on three main activities: transposing onto digital format the two parts of the dictionary, as well as its sources, correcting the digital format of eDA and eDLR, and building a register of software programs which will offer browsing capabilities, including direct access from the dictionary examples onto the pages of the original sources. This means that, besides all kind of browsing capabilities usual in electronic dictionaries, the user will also be able to click on an example and to obtain the view of a segment of the page in the original document from where the example was extracted (analogue with Google Books[4]).

At this phase there is no intention to acquire uniformity within the two parts of the digital dictionary, built very distantly in time, nor to correct and fill the gaps in DA, supposed to reflect with less accuracy the changes in the modern language (all entries belonging to letters which were left unchanged for more than 50 years). It is hoped that the process of updating the old parts of the dictionary to be made a lot easier by the existence of the electronic version.

There are many ways in which eDTLR, as a large dictionary built over two distinct periods in time, could be taken as a creative example for developing lexicographic resources of this type, in general, not only for Romanian. For instance, the vast collection of texts/attestations used to exemplify words and senses of the newer series of DLR (approximately 1,300,000 examples, representing about 88% of the whole text) can be used as source for updating the articles of the entries belonging to the old series of the dictionary, which, as said, do not reflect any more the modern language.

Then, we see eDTLR as opening the only thinkable way for a continuous process of keeping updated the dictionary thesaurus of a language, in the rhythm in which the language, a vivid entity, receives and accommodates new terms and senses, and forgets (and marks as such) obsolete terms and senses. Indeed, the society advances towards a status in which most, if not all, of the textual resources of a language will have an electronic copy[5]. Lemmatisation, part-of-speech tagging and detection of senses procedures have already become common components of the nowadays language technology. So, it is foreseeable a moment when the technology and the computing power will reach a level which will permit a

continuous processing of the huge collection of written materials appearing in one language. The least output that can be imagined with such a linguistic processing power-plant is the discovery of new words and senses entered in language, to be forwarded to the lexicographers for validation and inclusion in the continuously updated electronic dictionary of that language. Then, total lack or very sparse mentioning of a word or of a sense of a word for some time could signal that the word/sense became obsolete and, again, this has to be considered by the lexicographers.

Moreover, benefits of such a large digital dictionary go towards computational morphology of the language (for the exhaustive completion of the computational morphology in both analysis and generation), as well as towards the continuous enhancement of statistical-based language models. In computational semantics, such a dictionary, due to its richness in sentences exemplifying word senses, fills-up a tremendous need for a sense-annotated corpus, to be used for training a word sense disambiguation program, with applications in machine translation, information extraction, automatic detection of semantic roles of verbs and nouns derived from verbs.

The dictionary can be published cheaper by electronic means, while also providing sophisticated indexes between word occurrences, including links to occurrences outside the dictionary itself, in other linguistic thesauri or in other languages.

3. The Collaborative Approach in eDTLR

In order to reduce the expected time of proofreading necessary for professional lexicographers, we designed, implemented and advertised a Web-portal[6] with an editing window dedicated for corrections. As the Academy imposed restrictions on the dissemination of preliminary versions of the dictionary, for prestige and intellectual property rights (IPR) reasons, we had to find a compromise between our needs for accuracy and the perspective to involve a large community of volunteering proofreaders. The solution was to allow users to access only small extracts[7] during editing. The text displayed is under the limit of the IPR reproduction, and is assigned randomly every time the user asks for a new segment. When saving the processed document, it is integrated again in the whole. This strategy prevents users to re-assemble large portions of the dictionary. As the total amount of extracts reaches nearly 140,000, a rough estimation concerning the probability to obtain a given page from 12 extracts is of the order of 10^{-55}.

In order to motivate people we, decided to 'reward' contributors on the basis of the quantity and quality of their work. The 'reward' consists in advertising the best ranked volunteers and, eventually, on providing access to (parts of) the final product, once the project has reached its end. The problem that remains is how to evaluate the

---

[4] See www.books.google.fr, for instance.

[5] As felt in the research programs recently initiated (see in Europe CLARIN, for instance), but also in legislative initiatives promoting the recording of all publications for research.

[6] https://consilr.info.uaic.ro/edtlr/

[7] 10 – 12 lines on each column

quantity and quality of the work, and how to raise the level of accuracy.

3.1 How to Raise Peoples' Interest?

Collaborative projects like Linux and Wikipedia have always attracted many contributors because of the inherent intellectual challenge they pose to the volunteers. In the field of collaborative computational lexicography the experience has been sometimes promising, as in the case of Wiktionary, but sometimes showed a relatively little feedback from the public, as in the case with projects like Papillon, Kamusi.

The main type of entry in an encyclopaedia is the article describing notions, concepts, facts or events. The situation is different in the case of dictionaries. A dictionary is basically a set of data (definition, translation, grammatical information, related word) associated to a headword. The articles in an encyclopaedia give the author a great deal of freedom with respect to what s/he would like to focus on, what to include, at what level of detail, etc. Hence, the writer has a lot of liberty, which is not the case for the contributors of a dictionary, where the type of information to be contributed is decided beforehand by lexicographers. A dictionary entry is very rigid in terms of format and content and, usually, there are great academic debates on which words and which variants to include concerning the various senses, what a definition should look like, which specific examples to include (especially in the case of monolingual dictionaries), etc. Of course, all this looks more like a Procrustean bed than a creative activity, likely to motivate people.

The people of the Papillon project were painfully aware of this bottleneck. Actually, in order to solve it, a proposal has been made to convert the dictionary into a drill tutor or exercise generator, that is, a goal-driven, template-based sentence generator. The idea was to motivate people to contribute to the data base, by generating sentences based on their contributions (sentence patterns). Unfortunately, it is still premature to evaluate the heuristic value of this solution as the tool is still under development (Zock and Afantenos, 2007).

The specific framework in which we use volunteer work in eDTLR makes the whole enterprise even more dangerous (keener to rejection). The only implication of the user in our case is to ask her/him to improve the quality (i.e. correct) the output obtained from the OCR process, which is not really a very creative job. Actually, the corrector is supposed to spot and correct errors in the text in an editing window, provided on the right hand side of the screen, by comparing its content with the graphical image of the same segment of text, given at the left side of the screen (see Figure 1). So, how can we raise peoples' interest for this limited and, and not very enticing task?

The most important clue remains personal motivation, which we hope to raise in a large number of people, to contribute to a project which has a tremendous importance for the Romanian language. Calls have been spread over a diversity of channels, including mass-media and Internet, but also with the occasion of different scientific academic events. University professors have warmly embraced our objectives, starting to disseminate the eDTLR objectives and to persuade their students to

contribute. Moreover, the project consortium decided to stimulate contributors, based on an evaluation of the quantity and quality of their work. The stimuli consists of advertising the best ranked contributors and, eventually, on providing access to (parts of) the final product, once this is finalized[8]. The remaining problem is to recognise who are the people that deserve this distinction, therefore to evaluate the quantity and quality of the work. The task is not simple because counting only the number of sequences sent by each participant could encourage bad practice, as for instance clicking the Save button without any correction done, or typing blindly (therefore rather destroying the material than improving it). We think that an interface which feeds back to the user in a sensible and correct way, by producing encouraging or thanksgiving messages in cases of good practice and advertisements, although expressed in gentle phrases, in cases of bad quality, or even capable to totally block the access in cases of intentionally malefic interventions, can contribute in a substantial manner to the raise of the quality. This presupposes the ability to appreciate, as sensible as possible, the quality of the volunteered work. This issue is discussed in the following section.
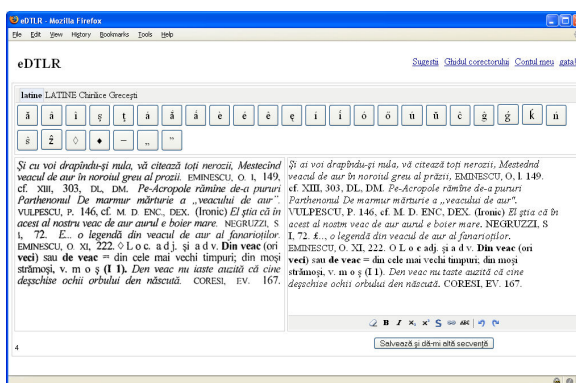

Figure 1. Screenshot of the on-line corrector interface dedicated to novice users.

3.2 How to Evaluate the Contributed Work?

We use manual and automatic procedures to evaluate the contributors' work. The manual evaluation is done by expert lexicographers during a second round of proofreading (whose main aim still remains quality enhancement). The portal allows the expert, recognised during login, to "follow basically the traces" of the same volunteer proofreader. By receiving the same sequence of screens as the novice, the expert's view concerning the anonymous contributor can stabilise, and, once having formed an opinion concerning the contributor, s/he will

---

[8] The consortium has not arrived yet at a consensus with respect to the channels on which eDTLR will be distributed (public on Internet, access based on subscription to different functionalities, DVD, etc.). The promises that we make now to our volunteer contributors refers to the case in which the final product will not be offered freely and completely on Internet.

have to rank the amateur contributor on a scale (ranging from excellent to malign) by using a facility offered by the interface.

Next, the contributor's record includes also an automatically computed evaluation based on the following criteria and heuristics:

- if a screen is saved without any key stroke during the editing phase, this is probably due to a lack of attention and should be ignored. As a result, the user will be demoted;
- if the edit distance for the input segment (as given by the OCR) and the output saved by the user exceeds a certain threshold [9], than the correction is unreliable and should be ignored. Again, the user is demoted;
- if there is next to no difference between the results of the expert and the prior results of a non expert, the non-expert will be promoted.

3.3 How to Obtain Accuracy?

Unlike Wikipedia, which tolerates a certain level of noise or incorrect material, the work we describe has to comply with the rigor and strict rules of the Academy, hence it is incompatible with errors. There are several ways in which to deal with accuracy.

First and foremost, the whole material of the dictionary will be corrected three times in a cascaded approach, meaning that the second and third correction phases are applied over previously corrected versions and not over the original. As shown already, the first phase is performed by (anonymous) contributors, while the remaining are executed by expert lexicographers. Secondly, a large source of errors, which come from the abbreviations of the bibliographical sources, is reduced in post-OCR processing. The dictionary has approximately 1.3 million citations, included to illustrate the use of word senses. For each citation a bibliographical reference is given. Having the full list of bibliographical references allowed us to identify the majority of them in the text using approximate string matching [10] and confidence values from OCR tool. Third, the interface has an ergonomic design, which, among others, allows zooming into different zones of the scanned image, in order to bring closer to the corrector's eyes portions otherwise difficult to read or to understand. Fourth, the user has the possibility to mark certain zones of characters as uncertain, attracting thus the lexicographer's attention to them for the next correction phase.

4. Practicing the Collaborative Approach

For the time being we have released a prototype of the online interface, which has been advertised among students in Computer Science. During a period of 6 months, in which we issued two calls, 119 students, out of a population of 1200, registered. From these, 35 students completed a single correction extract, 49 performed 2 - 5 corrections, 9 from 6 - 10, 15 from 11 - 30, 8 from 31 to 130 and 3 – more than 130 extracts. An exceptionally dedicated student proofread 312 screens for a period of two months. Overall, the effort amounts to 114 corrected dictionary pages contributed by less than 20% of the registered students. Although we expected more, the resulting profile corresponds to our Computer Science students, who tend to be more interested in evaluating the online interface than to help with the proofreading. Two out of the 119 students were easily identified as ill intended by simple heuristics.

The main features affecting the speed of correction were the extraction size, the OCR error rate and the ergonomics of the editing interface. With an extract of 10 to 12 lines, the correction time of a screen ranges between 30-120 seconds, with an average of 92 seconds.

Based on these counts, we estimated the effort and the total number of people required to accomplish the whole task of correction of the first phase. The total correction time estimated amounts to 3,577 hours, therefore 447 days, 8 working hours each (140,000 segments * 92 seconds / 3,600 sec/h / 8). Hence, if the average correction effort noticed during the first 6 month of the experimental setting is kept constant, the first phase will be finished within less than two years, conforming to the plan. Concerning the second estimation we considered a time frame of two years, estimating as 2,959 the number of collaborators to be involved, supposed to work at the rhythm observed (if 119 people have corrected 114 pages in 6 months, then 11,836 people are needed to correct 11,339 pages also in 6 months. Hence, only 2,959 people are needed to correct 11,339 pages in 24 months).

Both estimates are optimistic. The time estimate is clearly below the interval limit of the project. Finding about 3,000 people willing to work on this task will not be easy. Bear in mind though that the initial setting was a community of students in Computer Science, with preoccupations rather remote from computational lexicography. Moreover, the students were selected from a single faculty of one university in a country speaking Romanian. We expect that the invitation to contribute addressed to all categories of students over a larger territory will receive a much higher participation rate.

We investigate the idea of using free of errors text to improve the OCR rate of success in a continuous way. The volumes printed since nineties have been computer edited. As such, a small part of the dictionary content exists in electronic form free from errors and, therefore, will not need any correction. This material can be used to train OCR programs.

Another way to improve the OCR accuracy is by using an iterative process. In the current implementation, OCR processing, page splitting into extracts and randomization of user access to extracts are performed in this sequence, once for the whole text. Starting with a small amount of validated extracts, therefore as issued by the expert correctors, the process could be iterated, training the OCR engine and thus reducing the error rate in the next steps.

---

[9] The threshold is continuously updated based on extracts verified by experts for each dictionary volume.

[10] http://www.dcc.uchile.cl/~gnavarro/software

The number of extracts processed in a step will follow an exponential growth rate. For instance, at the first step, 8 pages could be chosen randomly, page splitting will follow and randomization of the obtained extracts. In the next step 16 pages will be processed, then 32, so on. The OCR training will happen on parallel, before the users actually consume the currently extracts under processing. The training cycle will be triggered by an alarm, which is chosen by taking into consideration estimations of correction time on one side and training plus processing time on the other side. Training will not start without the validation of experts for the current corrected extracts. Training on a very large corpus is not feasible and will be stopped when the accuracy will not improve significantly. Then, the processing of the remaining text could be done in a single sequence.

A different approach is that used in the Recaptha project[11]

5.   Conclusion

The paper studies different aspects related to collaborative approaches dedicated to lexicography, in the context of a project aiming to build one of the biggest digital dictionaries in the world. The setting is that of acquiring accurate data after scanning and processing by an OCR tool. The aspects we focus on are of great interest in a context where the acquisition of a very large, yet extremely reliable collection of lexicographic data is at stake at affordable costs. The question is how to discourage dissemination of unaccomplished or unreliable lexicographic material, while attracting a large community of volunteer contributors. There are also the problems of accuracy within a given setting, the problem of having many people with various backgrounds working together; how to motivate the largest number of potential reviewers; why it is important to evaluate contributors and how to do that.

Our proposed set of heuristics has been partly validated in an implementation. Based on the observation of the behaviour of the system over a period of 6 months, we were able to foresee the evolution of the enterprise until its final accomplishment. Contrary to other collaborative initiatives in lexicography, with unlimited perspectives concerning data acquisition, the scope of our work is clearly limited as is confined only to quality checking (proofreading) of dictionary entries. We give precise estimates showing that a collaborative approach could be a success despite the fact that the job in itself is not really very enticing.

6.   Acknowledgements

7.   References

*** (1965) Dictionary of Romanian Language. New Series.Tome VI. Romanian Academy.

Zimmer, B. (2007) Charting the Digital Future of Dictionary Research: Prospects for Online Collaborative Lexicography, communication at *DSNA* Chicago.

Zock, M, Afantenos S., (2007). Let's get the student into the driver's seat. *The Seventh International Symposium on Natural Language Processing*, Pattaya, Thailand.

---

[11] http://recaptcha.net/