Crowdsourcing terminology: harnessing the potential of translator’s glossaries

Ivanka Rajh*, Siniša Runjaić†

*Zagreb School of Economics and Management
Jordanovac 110, HR-10000Zagreb
ivanka.rajh@zsem.hr

†Department of General Linguistics, Institute of Croatian Language and Linguistics
Ulica Republike Austrije 16, HR-10000Zagreb
srunjaiic@ihjj.hr

Abstract

This paper describes the history of relations, separate processes of identifying user needs and connectedness of the results of two separate surveys, a convergence of positions and the ultimate establishment of cooperation between members of the Translators and Interpreters Interest Group (TIIG) and experts from the Department of General Linguistics at the Institute of Croatian Language and Linguistics. We’ll show how, regardless of the initial disparate positions, by forming smaller working groups, by accepting the idea of crowdsourcing and wider understanding of what terminological infrastructure means for two different types of stakeholders, a common goal can be defined leading to a relatively quick creation of a searchable database of translator’s glossaries. The emphasis is on the fact that such a database becomes an important resource for the translator community, but at the same time provides an important contribution to the increase in the number of domains and to the quality of results on the national terminological portal created within a scientific institution.

1. Introduction

When talking about specialized languages and people who use them on everyday basis, there is always a certain disbalance present and even emphasized between linguistic theories, or theories of terminology, and usability of applied terminological resources (databases, online dictionaries, glossaries) created on the basis of theories. Although terminology is a relatively young interdisciplinary field, which emerged in the second part of the 20th century accompanied by a mass of knowledge from its “mother” lexicography and by the advances in technology upon which its application can be based, even in the most recent literature we find conclusions such as this: “[…] translation of terminological theories into real and working terminographical products has so far left a lot to be desired…” (Fuertes-Olivera and Tarp, 2014: 128).

First discussions between today’s collaborators, members of the Translators and Interpreters Interest Group (TIIG) and terminologists from the Department of General Linguistics at the Institute of Croatian Language and Linguistics, as well as answers of the participants in the survey (Gracin et al. 2016), revealed a lack of trust of the users from the translator community in the applicability of the terminological resources built within the scientific community, which will be examined further in the following chapters.

The basic goal of this paper is to show how initial obstacles were overcome, the role that separate user surveys played in that, the current situation and methods of gathering material for the terminological database of translators’ glossaries and, finally, how that database is linked in a specific format to the metasearch engine within the Croatian Terminology Portal.

2. Translators as users of terminological resources

Translation professionals in Croatia have for a long time been organized into separate organizations depending on their specialization, such as the Croatian Association of Scientific and Technical Translators founded in 1957, The Croatian Society of Conference Interpreters (1974) and the Association of Court Interpreters and Translators (1989). Although the idea of an umbrella organization that would represent and promote interests of the entire translator community of Croatia was discussed on several occasions since Croatia gained its independence in 1990, it was only in 2009 that the Translators Group was formed within the Foreign Languages Association of the Croatian Chamber of Economy. The Translators Group was very active since the very beginning organizing its activities into eight special-interest areas, such as translation technologies, professional status and certification, and hosted several experts who held talks on topics related to translation, one of them being prof. Maja Bratanić from the Institute of Croatian Language and Linguistics, who presented the national terminology project Struna1 at the beginning of 2010. Terminology was quickly recognized as one of the most important issues common to translators of all specializations, and the members of the Group participated in the translation into English and German of university degrees, during which they established a close cooperation with the Agency for Science and Higher Education (Pavuna, 2011).

Eventually, in 2013, the Translators Group dissociated itself from the Foreign Languages Association and became an autonomous interest group within the Croatian Chamber of Economy, i.e. Translators and Interpreters Interest Group (TIIG). One of the aims of TIIG is a „creation of a comprehensive database which would serve translators in their work, containing sources such as dictionaries, style...
guides, subject-specific glossaries, thesauruses, web forum, collaborative, translation memories and special software...” (Pavuna, 2016). In order to fulfill this aim, a working group named National Terminological Infrastructure was set up in 2015, its main idea being to support translators in their efforts to exert influence on the creation of terminological infrastructure and adapt that infrastructure to their practical needs. Already at that stage, members of the working group were aware of the complexity of the project, which would involve not only translators, but also IT specialists, and planned to apply for the EU funds for the purpose of its financing as well as to internationalize the project thought cooperation with experts from Slovenia and TermNet. The first task of the working group was to assess the current situation and needs of the translators’ community. The following paragraphs describe the main conclusions of the survey conducted among the members of TIIG in early 2016 (Gracin et al., 2016). The survey collected answers to 37 questions from 99 translators within a period of around 40 days. The respondents were mainly freelancers (59.6%), followed by those employed in companies (27.3%), and the rest were part-time translators.

The survey showed that as many as 72% of translators find inconsistent, inappropriate or outdated solutions in terminological resources they use. When it comes to frequency of using particular resources, translators most commonly use the internet (73%), including online scientific and professional literature (65%), and multilingual texts (51%), followed by their own termbases and translation memories (40%), and inquiries among colleagues (32%), while print dictionaries represent the least consulted resource (21%). Taking into account that Croatian belongs to a group of non-dominant languages which create terms based on neology (Cabrè, 1999:18), it is not surprising that translators sometimes (49.5%) or even often (18.2%) have to create neologisms. However, 82% of them think creation of neologisms is not the only solution, probably referring to the fact that terminology exists but is not readily available. This assumption is corroborated by the fact that as many as 57.5% of translators use between 20% and 40% of their time on terminological research.

Although 78% of translators cooperate with their clients in their search for best terminological solutions, it seems that the clients are not aware of the importance and value of terminological resources. Namely, clients rarely or never (63%) provide them with terminological databases relevant for the translation project, and 82% of them rarely or never ask for a submission of the database created during the project. This means that a wealth of terminological data created in a time-consuming process remains underexploited, which points to huge inefficiencies of the current (non-existent) system of terminology management. The necessity of setting up a centralized terminological system or a database is clear taking into account that 92% of respondents do not know about all terminological resources available on the internet. When it comes to two most significant online terminological resources for Croatian, the Croatian Terminology Portal and IATE, the results are equally worrying: as many as 42% of translators are not aware of their existence, and roughly 70% of those who are informed consider that those resources do not satisfy the needs of translators. Those who do use the Croatian Terminology Portal2 are positive about it being reliable, but point to the fact that it contains a limited number of terminologies, that equivalents are available in only one or two major foreign languages, and that there are no colloquialisms and context, which translators find particularly useful. Respondents who use IATE confirm its exhaustiveness when it comes to the EU terminology, but complain about the limited number of Croatian terms, which are often inconsistent or downright wrong. It is important to emphasize that the quality of Croatian IATE has significantly improved since the beginning of 2016 when this survey among translators was conducted. According to the Croatian terminological team, the number of Croatian terms in IATE doubled from 10,005 to 20,526 between July 2015 and December 2017 (Miloš and Cimeša, 2017). Furthermore, Croatian IATE contains a low number of duplicates (only 2.5%) compared to “old” EU languages (11%), probably thanks to the inclusion of the Croatian team at the stage when system was already well functioning.

Finally, the survey shows that translators are willing to join forces in order to rectify this market failure, i.e. the lack of readily available, comprehensive and reliable terminological resources. The majority of respondents (78.8%) have their own terminological databases, and 65.7% share those with their colleagues. Furthermore, as many as 96% of translators would use a centralized terminological database and 77.8% would help in its creation, and even pay for its maintenance (65.7%). Additionally, translators consider that terminological databases, just like any knowledge, should be available to everyone regardless of their financial status. Some suggest that such database should be created and financed from the state budget and thus available to everyone for free or at a small fee like public libraries.

The results of the survey led to the overall conclusion that Croatian translators are not satisfied with the current state of terminological resources for Croatian language, especially by their suitability to translators’ needs, their volume and quality. On the other hand, it was also revealed that many translators are not aware of the existence of different terminological resources for Croatian. Therefore, the working group suggested that TIIG should engage in the creation of a database of translator’s glossaries, which would become an element of a web portal containing a comprehensive overview of information necessary for high quality translation work, including links to language resources, translation repositories and IT tools for translators.

3. Terminological infrastructure from the Institute of Croatian Language and Linguistics

This chapter shall clarify how, in the scientific community, the basis for the creation of the terminological infrastructure, Struna and the Croatian Terminology Portal was established, making it logical for TIIG, i.e. TIIG's
working group National Terminological Infrastructure, to choose for their endeavor as partners and collaborators precisely those resources and their administrators. The attempts of establishing publicly-accessible infrastructure for specialized languages within the framework of the standard Croatian language were made on several occasions during the second half of the 20th century, when certain committees for creation of terminological dictionaries were set up, but it was only within the process of preparation for the accession of the Republic of Croatia to the EU, which included a major project of the translation of the aquis communautaire and accompanying documents, that a more organized terminological work on the national level began. In parallel to those processes, the scientific community realized that the establishment of a modern terminological database is a necessity, so the Croatian Science Foundation first financed a project within the program Sociocultural transition from industrial into the knowledge society, designating the Institute for Croatian Language and Linguistics as a national coordinator for the creation of Struna, a database of Croatian special-field terminology (Brač, Bratanić and Ostroški Anić, 2015: 10–15). Struna was founded in 2008 and represents a traditionally-organized normative terminological database, which is populated through a special system of scientific and professional projects financed by the Croatian Science Foundation. The projects are implemented by experts in different scientific fields assisted by linguists, i.e. terminologists-terminographers (Bratanić and Ostroški Anić, 2015: 59–68). Struna currently encompasses 20 finalized projects covering diverse fields in humanities, interdisciplinary, natural and applied sciences. Since February 2012, the general public can access more than 30,000 concepts, i.e. standardized, preferred terms, accompanied by different synonyms (admitted, deprecated, obsolete and jargon terms). Struna can serve translators as a stable source of information since every term must have at least an English equivalent, but often there are also equivalents in other European languages such as German, French or Russian. Due to the complexity of data in Struna, the search engine includes both simple and advanced types of search, with the help of special characters (wildcards) and Boolean operators. The Croatian Terminology Portal was conceived primarily as a user- and translator-oriented addition to the Struna system, as well as a central place of gathering diverse terminologies and it was publicly released after two years of preparatory work in July 2015. The portal search engine was designed as a metasearch engine, or aggregate search engine, and is simultaneously searching four separate resources: Struna, donated terminological dictionaries transformed into a terminological database form, digitally accessible resources of the Miroslav Krleža Institute of Lexicography and terminology collections of the Croatian Standards Institute (Bratanić, Ostroški Anić and Runjačić, 2017: 663–64). Since terminologies in the portal system are typologically more diverse than the Struna database, built according to the ISO recommendations and in the TBX format for data exchange, the search engine of the portal was completely simplified and made to simultaneously search all available terms in any language for a string of characters entered into a search bar, and the results can be further narrowed down and sorted depending on the user needs and preferences. There are more than 100,000 Croatian terms and more than 160,000 equivalents in foreign languages available in the portal, which contains a much wider range of terminologies than Struna. While Struna gained recognition among users as a reliable source of information since its public release in 2012, a new strategy had to be devised for the promotion of the Croatian Terminology Portal after it was launched in 2015. Since the Portal was made simple to use with the translator community in mind, the information on its launch was sent to all translator associations and translation agencies. Among the first to react was the Translators and Interpreters Interest Group (TIIG) of the Croatian Chamber of Economy, which invited authors to present the Portal during the annual meeting at the beginning of 2016. From then on, the idea of the creation of the database made up of translator’s glossaries was gradually developed, including the idea of its inclusion into the Croatian Terminology Portal.

At the same time, a survey on needs and habits of the users of both terminological resources maintained by the Department of General Linguistics of the Institute of Croatian Language and Linguistics was carried out. The results of the survey were presented at the scientific conference Slavic Terminology Today in Belgrade in May 2016 (Lončar and Runjačić, 2016). For the purpose of this paper, we are going to focus on those results of the survey which correspond to the results obtained by the Translators and Interpreters Interest Group (TIIG) in their survey, and which prove that the cooperation between two organizations is a logical outcome of the answers provided by the respondents of survey conducted by the Institute³. The survey was made up of 32 questions, completed by 85 respondents. Demographic data point to the connection with the translators’ survey due to the similar age and educational structure of respondents. Namely, the majority of respondents were in the 25-46 year age group (70.6%), of which 90.6% with a university-level diploma, while 79.8% said they were translators of different legal statuses. The most interesting information was that 63.4% of respondents use electronic dictionaries and terminological databases on (almost) daily basis, while only 20.2% search printed terminological resources. At the time, the Croatian Terminology Portal was open for public for less than 365 days, so the most important question was the one regarding frequency of visits to the Portal. The results showed that most respondents search Struna several times a month (71.8%), and 7.1% every day, while the Croatian Terminology Portal was still not as recognized in the public since only 46.3% respondents were searching it several times a month. Therefore, it was considered that the cooperation with the Translators and Interpreters Interest Group (TIIG), and especially the inclusion of translator’s

³ Of course, since the survey was taken by anonymous respondents, we cannot know whether the members of TIIG also took this survey.
glossaries alongside other resources as an integral part of the Portal search engine could significantly improve the number of visits by translators, thus improving its visibility with the aim of fulfilling its primary purpose on the one hand, and reducing the possible dissatisfaction of translators by an apparently limited scope of terminological resources publicly available in Croatia on the other.

4. Construction of the database and the search engine

Technical features and possibilities of search on the Croatian Terminology Portal were presented to the TIIG in April 2016, with a special emphasis on the new database “Terminology dictionaries and glossaries” available on the Portal. Unlike Struna, which is based on as many as 46 fields in accordance with the ISO terminology standards and the TBX (Term Base eXchange) standard and set up for a specific type of term processing in cooperation between experts and terminologists (Brač and Lončar, 2012: 261–66), the new database is significantly simplified in structure and adapted for the import of ready-made terminology manuals, which needed only minor adjustments and transformation into a terminological database form. That database is made up of basic metadata for individual terminological resource and the minimum terminological entry consists of at least one Croatian term and an equivalent in one of the European languages, depending on the type of document being converted. Thus, the database comprises both simple bilingual dictionaries without definitions4 and, for instance, quadrilingual lexicons with longer definitions and notes, of course with the authors’ permissions and respect of copyright5.

Members of the Council of the Translators and Interpreters Interest Group (TIIG) recognized that such a system of a simplified database corresponds to their idea of terminological infrastructure for translators comprised of translator’s glossaries (based on the crowdsourcing principle), and from that moment on, negotiations and preparations for the creation of the database and the search engine for translator’s glossaries of the TIIG took off. It was arranged that the experts from the Department of General Linguistics, as the authors of the resource and active terminologists in the Struna and the Croatian Terminology Portal projects, provide their technical and expert know-how and experience in building the system, while the members of the Council of TIIG will invite interested members to prepare and send their glossaries to be imported into the database. The agreement was finalized at the beginning of 2017, when common efforts were invested in designing the final version of the database of translator’s glossaries. We can conclude that, during the creation of the database, the potential of scientifically-based terminological specifications and the internal content management system (CMS) was used, so today we use a simple procedure of importing glossaries in their classic table formats (.xls, .xlsx, etc.), which can subsequently be additionally edited by administrators in the database itself. The final searchable entry always contains the data relevant to translators such as the name and topic of a glossary, field classification, Croatian term (and possible synonyms), equivalents in foreign languages (depending on the language specialization of the translator who is the author of the glossary), and additional fields for possible information on the context and any additional remarks.

Figure 1. An illustration of the record radna snaga ‘labour force’ in the search engine.

The searchable database was released to public in April 2018 on the website of the Croatian Chamber of Economy. The condition of the experts from the Institute of Croatian Language and Linguistics was for the database to be hosted on the same server as the other terminological resources. Furthermore, the TIIG database was linked through the application programming interface (API), thus becoming searchable on the Croatian Terminology Portal in July 2018 for further improvement of visibility and quality of results for all interested publics. Currently, it contains 19 glossaries with 11,124 Croatian terms and 14,854 equivalents in different foreign languages, and there are a few more glossaries in the pipeline. Its simple query system is in line with the needs of translators for simple term searches of any string of characters in any language6.

5. Features of translator’s glossaries

As expected, glossaries that have been sent for the inclusion in the TIIG’s terminological database vary in structure, length and overall quality, so they require additional editing performed by the terminologists from the Institute. When invited to send their contributions, translators were given very broad guidelines as to the structure of their glossaries, which were further discussed during the workshop organized for that purpose in the Croatian Chamber of Economy in July 2017. An important aspect of that well-attended workshop was to remove any doubts and fears translators might have had regarding their glossaries and explain the overall benefit of their

---

6 Which also corresponds to the results of the survey (Lončar and Runjać, 2016) according to which 2/3 of the users of Struna and the Croatian Terminology Portal prefer simple search as opposed to advanced search with the help of Boolean operators.
publication for the authors themselves and the entire translator community. As to the guidelines, they were the following: glossaries should preferably be in a table format, contain a minimum of two columns (terms in Croatian and a foreign language), be consistent in the use of different symbols (e.g. semicolons to separate synonyms), cells should contain one piece of information, terms should be written in consistence with the language rules, the information on gender may be omitted. Translators were encouraged to send even short glossaries of as few as 20 terms and supply the name of the glossary, subject field and the author’s name even though the authors may request to remain anonymous to the public.

The database was made very robust so as to accommodate different formats of glossaries. The content management system allows import of different categories of data (term, name of the glossary, subject field, synonym, context, source, remarks etc.), of which only four are mandatory, i.e. the term in Croatian, its equivalent in a foreign language, the title of the glossary and the subject field.

The glossaries received until now indeed vary in terms of their volume and structure: from a big Croatian-English-Slovene central banking glossary of 1720 entries with definitions to a small Croatian-Italian inheritance law glossary with 51 pairs of equivalents. Other subject fields covered by the received glossaries are: insurance, human bones, saltwater fish, winemaking, weather, real estate, world languages, scripts and regions, single euro payments area (SEPA), waters, and ecology. Editing performed on those glossaries included additional technical and orthographical corrections in the columns carried out by the Institute's members of the project, such as minor features that would not be compatible with the database CMS once the file is uploaded to the system (i.e. usage of brackets denoting the synonymy instead of two separate terms, missing or false diacritics in Croatian etc.).

Project coordinators are aware of the limitations of translator’s glossaries, which shall be clearly stated on the webpage of the terminological database. Translator’s glossaries are the result of a translation project and as such reflect the process of a terminology search translators find themselves in. Quite often, translators are faced with a lack of terminological resources for a particular subject or inadequacies of existing resources, so they are forced to compile their glossaries from a variety of sources of different quality and origin. Users of translator’s glossaries should bear in mind that they contain terms found in the text that was being translated and not the terminology of an entire subject field. Furthermore, they may contain terms that were specific to a particular project or client requirements. However, it is expected that terms that are found in the database, even though they may not be the ones they are looking for, may help skilled translators turn their search in the right direction.

In order to populate the database with as many glossaries as possible, TIIG plans to continue promotional activities among the translator community in Croatia via different channels (workshops, website, Facebook). Additionally, project members shall contact translation services of various government institutions with the aim of detecting internal terminology resources and persuading the owners to enable the access to the general public either by their inclusion into the translator’s glossary database or directly into the Croatian Terminology Portal. Namely, members of TIIG consider that terminological resources that were created by using taxpayers’ money should be made available to those who paid for them.

Additional confirmation of such a belief came from the Slovenian Jožef Stefan Institute, whose researchers Simon Krek and Andraž Repar contacted TIIG in May 2018, expressed their interest in the translator’s glossaries database and suggested cooperation on further collection of terminological resources, hoping that by common action, we could persuade the owners of such resources, especially public institutions, to offer them for public use. Ultimately, the collected terminological resources, including the translator’s glossaries database, would be used within the eTranslation Termbank project, co-financed by the European Union's Connecting Europe Facility and implemented by a consortium of partners, among which Jožef Stefan Institute, is in charge of collecting resources in Slovenian, Croatian and Bulgarian.

6. Conclusion
In previous chapters, we have described a specific example of collaboration in which two separate ideas on the establishment of a comprehensive terminological infrastructure are brought together in a common IT project. Although initially the need for the creation of a database of translator’s glossaries was expressed by a special-interest community, i.e. Translators and Interpreters Interest Group (TIIG), flexibility of the experts from the Department of General Linguistics of the Institute of Croatian Language and Linguistics and their technical capacities needed for the creation of an adapted terminological database provided an additional benefit for a wider translator community by inclusion of translator’s glossaries into the system of the Croatian Terminology Portal. We have also demonstrated the importance of conducting user surveys for further planning of implementing activities by both the TIIG and the Institute. The readiness of the participants in the surveys to accept the idea of crowdsourcing and open access to information as a necessary condition to enable the technical realization of the described process also proved to be of a great importance. The database of translator’s glossaries described above, its search engine and linkage with the search engine of the Croatian Terminology Portal has been

---

7 http://ettb.ijs.si/sl/etranslation-termbank/.
8 According to the integrated definition (Estellés-Arolas and González-Ladrón-de-Guevara, 2012): “Crowdsourcing is a type of participative online activity in which an individual, an institution, a nonprofit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task; of variable complexity and modularity, and; in which the crowd should participate, bringing their work, money, knowledge **[and/or]** experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and use to their advantage that which the user has brought to the venture, whose form will depend on the type of activity undertaken.”
made fully operational and open to public. The next step is to conduct surveys and similar activities in order to check the satisfaction of users with the performance and usefulness of the entire project, as well as to encourage them to participate in the creation of the database by contributing their glossaries to the project.

7. References


Denis Gracin, Tomislav Kojundžić, Vesna Kaniški, Ivanka Rajh and Damir Pavuna. 2016. Prevoditelji i terminologija. Survey conducted among the members of Translators and Interpreters Interest Group (TIIG) within the Croatian Chamber of Economy, Zagreb, from January 20 to March 8 2016.
