Harmonization of Multiple Entries in the Terminology Database *Struna* (Croatian Special Field Terminology)

Marina Bergovec, Siniša Runjaić

Institute of Croatian Language and Linguistics, Zagreb, Croatia

{mbergov, srunjaic}@ihjj.hr

Abstract. This paper aims to present our experience in analysing multiple entries during the term harmonization process in the Croatian terminology database *Struna*. The first part will demonstrate that a great number of multiple entries before harmonization were the result of: a) Its organizational structure, in which there is no direct cooperation between field experts from individual projects, and b) Inexperience of the members of individual projects in distinguishing between language for special purpose and general-purpose language. The need for harmonization became evident at the moment when merging of all terms merged into a single search engine. The harmonization was based on a list of multiple entries and began at term level. The central part of the paper will present the terminographical methods and theoretical background used in the process of distinguishing homonyms from synonyms, and in the process of synonym normalization.

Keywords. Croatian special field terminology, multiple entries, data harmonization, homonymy, synonymy

1 Introduction

Struna termbase is a prescriptive monolingual national terminology database. Its forming was initiated by the *Croatian Standard Language Council* in 2007 which encouraged the launch of the *Development of Croatian Special Field Terminology* (*Struna*) project. Financed by the *National Foundation for Science, Higher Education and Technological Development of the Republic of Croatia* it was envisaged to create the basis for the standardization of terminology in the Croatian language for all major professional domains. The *National Foundation for Science, Higher Education and Technological Development* holds a public competition once or twice a year, which is open to all institutions of higher education and experts from various subject domains (preferably experienced in terminology work). Linguistic and terminographical verification and technical support is provided by the *Institute of Croatian Language and Linguistics*. The selection of the domains is made according to the quality of individual proposals and the need or requirements for the terminology of certain domains. *Struna* presently includes the following domains: aviation and air traffic, cartography and geoinformatics, chemistry, civil engineering, corrosion, dental medicine, marine

Aguado de Cea et al. (Eds.): Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012), pp.231-241. 19-22 June 2012, Madrid, Spain

engineering, mechanical engineering, polymers, and physics. Maritime terminology, anthropological terminology, and EU law terminology will be available in the autumn of 2012. Also, new projects are about to commence, including medicine (anatomy and physiology), archaeology, mathematics, and hydraulics.

Struna was developed in the context of globalization and intensive re-evaluation of the position of national terminology as the alternative to English which has become the language of global communication. Therefore, considerable and constant effort is made to preserve and develop national terminology [1, p. 18]. The standardization of Croatian terminology includes all aspects of a complex process: the unification of concepts and concept systems, the definition of terms, the reduction of homonymy, the elimination of synonymy, the fixing of designations, including abbreviations and symbols, and the creation of new terms. [2, p. 200] It is both language-oriented and subject-oriented.

Struna's structure reflects an idealistic vision for a national prescriptive termbase focused on the sociolinguistic need for language and terminology planning in the Croatian standard language. Thus, the first version of the termbase interface had a simplified structure. Very soon, as it grew in its number of terms, it became apparent that language for special purpose requires more elaborated processing. LSP appears in natural communication, does not belong to artificial language and the methodology for its analysis has to be modified in order to be more descriptive. [3, p. 42] Therefore, the interface had to develop technically in order to be able to handle the complexities of LSP best as possible. Some fields and categories were modified and new ones introduced in accordance with the requirements of terminographical description.

2 Terminology Work in Croatia

Terminology in Croatia is a young discipline, although its beginnings can be traced back the 19th century. Bogoslav Šulek was a Croatian linguist and lexicographer who advocated linguistic purism and was highly active in the creation of neologisms. Although some of them were not recognized or accepted in his time, many of them have become part of the Croatian standard and the language for special purpose. Since his time, during the 20th century a considerable number of dictionaries of the language for special purpose for various domains have been published. The end of the 20th century saw significant and far-reaching changes in the history of the Croatian language. The Republic of Croatia declared its independence in 1991 and the re-standardization of the Croatian standard language began in 1992. In 2007, with *Croatian Standard Language Council*'s initiative launching the *Development of Croatian Special Field Terminology* project, a process of terminology re-standardization involving several national institutions and representative bodies commenced.

Terminologists involved in the process of terminology standardization are junior researchers and researchers from the *Institute of Croatian Language and Linguistics* whose terminological education began along with the project, with valuable help and professional guidance from senior colleagues experienced in lexicographical and terminographical work.

Taking these circumstances into consideration, *Struna* underwent a number of challenges at the outset of the project. The aim of this paper is to present a theoretical and terminographical framework and our practical experience in analyzing multiple entries during the harmonization process for concepts and terms in the *Struna* termbase.

3 The Data Harmonization Process

During the harmonization process, around 10 percent of approximately 16,000 database entries were isolated as homographs and potential synonyms. A detailed and systematic analysis showed that there were two main reasons for this relatively high percentage of multiple entries.

3.1 The Compilation of Terms

The compilation of terms for *Struna* differs from the work organization typical of national terminology centers. Joint efforts and coordination of subject field experts and terminologists is seen as the best model for the standardization of Croatian terminology, since subject specialists are the end-users. Each project is treated individually in its terminographical analysis, meaning that the terminology of each subject field is compiled and analyzed separately. The list of terms requiring terminographical analysis was drafted by field experts, however without the obligation to share it with the assigned terminologist. Consequently, this severely hampered terminologists' work, since the analysis did not start from concepts within a concept system but from terms themselves.

The need for harmonization became evident at the moment all the terms were merged into a single search engine. Since several projects were conceptually related (chemistry, corrosion, polymers, physics), there was a considerable number of shared concepts. Additional problems were sometimes generated by the field experts, who did not fully understand what harmonization meant and failed to accept the principles of harmonization that the terminologists followed [4].

3.2 The Indistinct Boundary between Language for Special Purpose and General-purpose Language

According to Cabré, „the purpose of terminological standardization is to aid communication in special languages, and is not applied to the vocabulary of the general language" [2, p. 200]. However, field experts quite often failed to distinguish between language for special purpose and general-purpose language. A significant number of general-purpose language items were seen as part of LSP. This naturally led to numerous multiple entries, since different, yet related domains (e. g. polymers, corrosion, and chemistry) defined the same concepts differently. In keeping with *Struna*'s initial aim to be prescriptive, all multiple entries required further analysis and harmonization.

Further terminographical and linguistic analysis was also necessary for preferred terms that differed either phonologically, morphologically or syntactically from the Croatian language standard[1]. In such situations, Croatian standard language specialists suggested a proposed term. However, such a term was not always acceptable to field experts, thus generating a significant amount of discussion between terminologists and standard language specialists.

4 Theoretical and Terminographical Background

During the process of harmonization of terms and concepts included in the Struna termbase before it was made available to the public, we presumed that not all of the homographs fell under the technical category „doublette“[2] and that the starting point of our analysis should be to separate homonyms (two or more terms that are spelt the same way but may justifiably be listed as separate entries as they denote slightly or completely different concepts) from synonyms (doublettes or multiple entries that are spelt the same way and that are in two separate entries and refer to the same concept). [6]

The theoretical background of our analysis was based on the standards defined in ISO 860:2007 (Terminology work – Harmonization of concepts and terms) and ISO 704:2009 (Terminology work – Principles and methods)[3]. Additionally, considering the fact that the initial version of the termbase was based on the traditional theory of terminology and a lexicographical approach, we were aware of the constant need to re-evaluate it, since it had become obvious during the term compilation process that not all terms are clear-cut cases of unique preferred terms. „Only when the compilation of terms is completed, during the process of analysis and problem-solving, can the appropriateness of an intervention to reduce any existent variant be considered“. [3, p. 42]

Bearing all this in mind, our main hypothesis was inspired by the socio-cognitive terminological work of R. Temmerman, with special reference to the part which deals with the questioning of the univocity ideal of traditional terminology. In other words,

---

1 „Current problems and challenges in term formation also include discrepancies with respect to general linguistic models in morphology, diversity and inconsistency of rules in different domains (in particular for natural sciences with specific nomenclatures), lack of detail in the description of many languages, and the need for full codification of these languages (e.g. through language planning) in order to have reliable rules for terminology development, in particular concerning orthography, spelling, pronunciation, and grammar“. [5, p. 11]

2 „According to ISO TR 26162, a doublette is a *terminological entry that describes the same concept as another entry*.“ [6]

3 „Incidences of homonymy and synonymy usually lead to the need for term harmonization, which is part of the standardization process. The standardization of terminologies in various subject fields frequently implies harmonization of concepts and/or terms within a subject field, across subject fields and across languages. To reduce duplication and to reduce the high cost of terminology work, efforts should be made to harmonize whenever minor differences exist. See ISO 860 for principles of term and concept harmonization.” [7, p. 35]

we presumed that a certain number of homographic pairs or groups can be attributed to the principle of univocity[4], but that a representatively larger number of multiple entries would include cases of synonyms[5] which have polysemy and metonymy as part of their naming history. [8, p. 67]

Following this, our projection was that the analysis of the second group of homographs belonging to the prototype structure of Temmerman's framework would be more time-consuming and would require more expert and terminographical knowledge during the harmonization process.

5 Analysis of Multiple Entries

The aim of our analysis was primarily to present to subject field experts how doublettes can deteriorate the quality of a terminological database, and to offer the most acceptable terminographical descriptions of multiple entries to experts from certain domains. Although the starting point of our analysis was a simple list of homographs extracted from the termbase, we were also quite familiar with the notion that there could be a wide range of possible conceptual doublettes hidden across domains.[6]

The analysis began in January of 2012 and was performed on a sample of 19,384 terms from the termbase, of which 1920, or 9.9 percent, were terms requiring further analysis and harmonization. Statistically, this included 860 multiple entries, of which 673 were doublettes, while the remaining 187 appeared in the termbase more than twice. During the process of intense terminographical work leading up to the opening of the *Struna* termbase to the public (February 21st, 2012), 151 multiple entries were normalized and removed from the initial list in collaboration with field experts. Final corrections in the terminology of physics were made during January and February, when 164 examples of multiple entries were discussed and also eliminated. Research for this paper was finally performed on a sample of 545 multiple entries still visible online in the search engine of the termbase.

5.1 The Analysis Process

According to the principles and methods elaborated upon ISO 860[7] and types of designations from ISO 704[8], we attempted to distinguish whether multiple entries belonged to the categories of homonymy or synonymy during the comparison of concepts and terms.

---

4 We expected mainly the terms with already defined superordinate concepts and their preferred term (*type of*, *part of*, *consists of*, etc.) to appear in this group.

5 We presumed that gerunds and other nouns with their superordinate concepts like *process*, *property*, *action*, *device*, etc. will appear in this group.

6 „An entry must be in line with its conceptual system – whether that system is made explicit in the database or not – as the terminology database is only sustainable as the ontology formed by these conceptual entries." [9, p. 130]

7 [4, p. 16]

8 [7, pp. 34–36]

The following cases were classified as homonyms:

Where preferred term A and preferred term B (or more) within a single project or shared between different projects define different concepts, or preferred term A and preferred term B (or more) define similar concepts, with specific meanings within the framework of the language for special purpose, but different enough to justify their existence as separate entries in the termbase.

The following cases were classified as synonyms:

Where preferred term A and preferred term B (or more) define and relate to similar concepts and their subject domains and specific meanings within the framework of the language for special purpose are almost identical. In such cases, the working group, together with field experts and other coordinating partners, proposed one of the following two methods:

(a) Method A – If, from a terminographical point of view, one of the entries encompasses the concept in a more satisfactory way than the other(s)[9], the working group proposes that only that entry remain visible to the public, and
(b) Method B – If, from a terminographical point of view, none of the entries encompasses the concept in a satisfactory manner, the working group takes all data from given entries into account and harmonizes the definition in order to consolidate them into a single entry in the termbase.

5.2 Preliminary Results of Analysis

After a thorough comparison of 535 multiple entries according to the given principles and methods, our preliminary results showed that 109 doublettes (20%) were homonymous pairs, while 436 multiple entries (80%) could be considered synonyms in need of harmonization (See Fig. 1.).

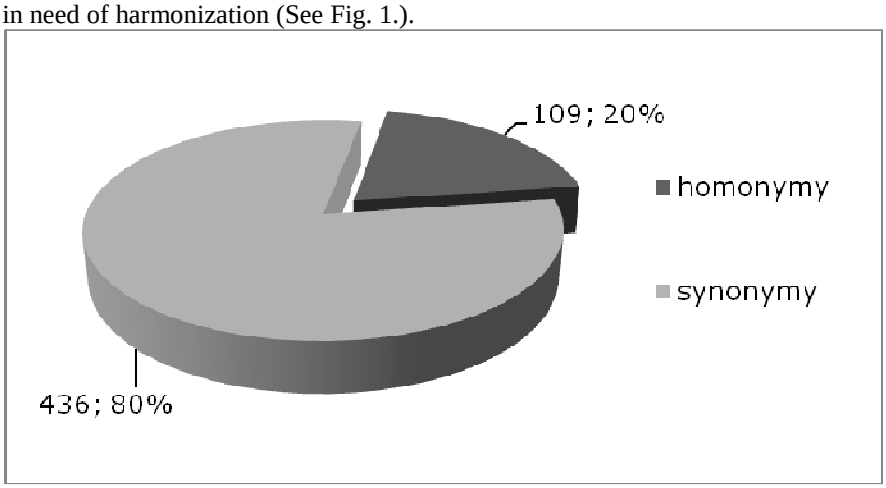

Fig. 1. The distribution of homonyms and synonyms (from the list of 545 multiple entries)

[9] The main criteria were more accurate definitions in describing given concepts according to ISO 704 [7, pp. 22–34] and lists of subordinate concepts.

Here are some of the illustrative examples of homographs that we considered homonyms.[10]

Table 1. Example 1 '*mjehur*'

mjehur	**mjehur**
bula	
<*oral medicine*> patološka promjena veća od 3 mm u promjeru, koja može dosegnuti i promjer od nekoliko centimetara	<*polymers*> uzdignuće površine različitih oblika i izmjera ispod koje se nalazi šupljina
engl. *bulla*	engl. *blister*

Table 2. Example 2 '*mjehurić*'

mjehurić	**mjehurić**
vezikula	
<*oral medicine*> eflorescencija nad razinom sluznice, veličine 1–3 mm, ispunjena seroznim ili hemoragičnim sadržajem	<*physics*> šupljina kuglastoga oblika koja je ispunjena plinom, a omeđena je tankim slojem fluida ili je uronjena u fluid
engl. *vesicle*	engl. *bubble*

Examples 1 and 2 show homographs '*mjehur*' and '*mjehurić*' which denote different concepts from very different domains.

Table 3. Example 3 '*vrijeme*'

vrijeme	**vrijeme**
<*geoinformatics*> jedna od triju sastavnica geoprostornih podataka, uz prostornu i tematsku	<*physics*> veličina koja obilježava trajanje zbivanja ili razmak između dvaju događaja
engl. *time*	engl. *time*

Example 3 '*vrijeme*' shows a slightly more complex type of homonymy where the same concept is defined, but where, in our opinion, both terms inherit different characteristics from their superordinate concepts ('*time*' as one of the three obligatory types of geoinformatical data and '*time*' as one of the major physical dimensions), allowing them to be considered homonyms due to the concept systems within their domains.

[10] These examples are simply illustrations of more complete terminological records in the Croatian language, and they include the preferred term, admitted term, domain, definition and foreign language equivalents for term A (left) and the term B (right).

Similarly, as in the examples shown above, we briefly described the remaining 106 multiple entries and suggested they remain separate entries in the search engine.

The next step was more time consuming and more demanding of the terminologists, since they had to choose the best methods to harmonize 436 multiple entries that were considered synonyms. The first results show that, in over 51 percent of cases (224), one definition between two or more could be chosen representative enough to denote the concept in question and acceptable to end-users, while for the remaining 49 percent of cases (212) it was considered more appropriate to consolidate various data from several entries into a single record to replace them in the termbase. (See Fig. 2.)

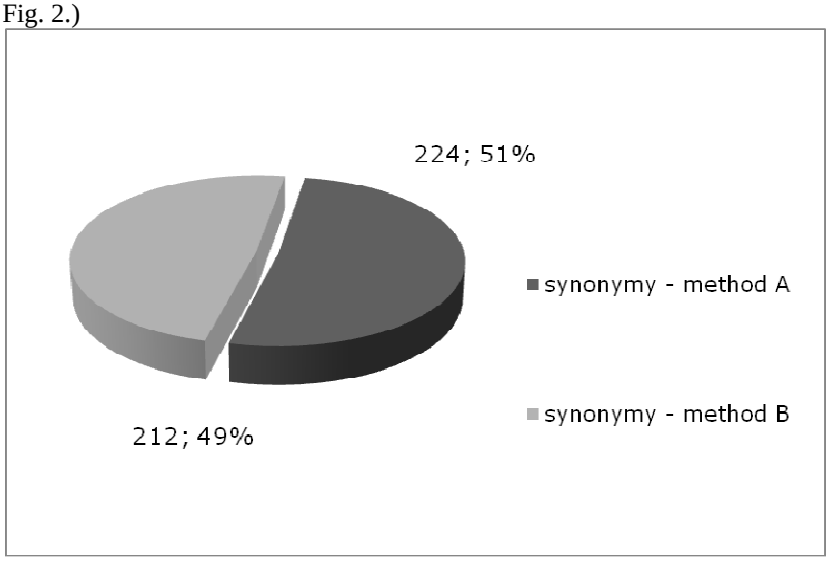

Fig. 2. The distribution of proposals for method A (choosing the most representative entry for end-users) or method B (consolidating entries into a single entry) from the list of 436 remaining multiple entries considered synonyms

Below are two indicative examples for method A. After the analysis of the respective concept systems within all the projects and domains in the database, we proposed to the domain experts that designations for terms A in Examples 4 and 5 should not be accepted as representative, but rather those for terms B as they define the concept more accurately and are more congruent with *Struna* terminographical principles and methods.

Table 4. Example 4 '*deoksiribonukleinska kiselina*'

deoksiribonukleinska kiselina DNA <*medicine*> nukleinska kiselina koja nosi genski materijal svih staničnih organizama engl. *deoxyribonucleic acid*	**deoksiribonukleinska kiselina** DNA <*chemistry*> nukleinska kiselina sastavljena od dvaju lanaca polimera deoksiribonukleotida međusobno uvijenih u obliku dvostruke spiralne zavojnice engl. *deoxyribonucleic acid*

Table 5. Example 5 '*flokulacija*'

flokulacija <*chemistry*> nastajanje pahuljastih nakupina koloida engl. *flocculation*	**flokulacija** <*chemical engineering*> izdvajanje koloidnih čestica nakupljanjem u nakupine čestica do veličine pri kojoj dolazi do taloženja engl. *flocculation*

The next group of examples are different types of multiple entries that, according to our analysis and for various terminographical reasons, needed to be harmonized using method B.

Table 6. Example 6 '*povećalo*'

povećalo <*physics*> konvergentna leća žarišne daljine manje od daljine jasnoga vida koja služi za gledanje bliskih sitnih predmeta engl. *magnifying glass, magnifier*	**povećalo** <*medicine*> konveksna leća za uvećanje slike sitnih predmeta na maloj udaljenosti engl. *magnifying glass, magnifier*

Definitions of designations '*povećalo*' in Example 6 begin with different superordinate concepts (type of lens), and other data in the termbase do not provide sufficient information as to which concept system it belongs to. In such cases, experts from the given domains should contact their terminologists and write a single harmonized definition of the preferred term.

Table 7. Example 7 '*adhezija*', '*prianjanje*'

adhezija <*physics*> međusobno privlačenje površina dvaju tijela koja su načinjena od različitih tvari engl. *adhesion*	**adhezija** prianjanje <*biology*> molekulska sila kojom se površine dvaju tijela međusobno privlače engl. *adhesion* **prianjanje** adhezija <*chemical engineering*> privlačenje između suhoga sloja premaznoga materijala i podloge na koju se nanosi engl. *adhesion*

Example 7 shows three entries that refer to an identical concept. Definitions of Term A and Term B '*adhezija*' could easily be consolidated into a single entry. The definition of Term C '*prianjanje*' refers to the same concept, but with a meaning specific to the field of chemical engineering. In addition, there are differences in attitude towards the normative status of international and Croatian designations of the same concept, so further harmonization is necessary.

Table 8. Example 8 '*anoda*'

anoda <*physics*> elektroda koja ima veći električni potencijal engl. *anode* **anoda** <*engineering*> pozitivna elektroda u elektrolitskome članku engl. *anode*	**anoda** <*chemistry*> negativno nabijena elektroda engl. *anode* **anoda** <*chemical engineering*> elektroda na kojoj prevladava anodna reakcija engl. *anode*

In Example 8 '*anoda*' it is obvious that all four projects aimed at defining the same concept. All four definitions are more than misleading because term B and term C begin with antonymical superordinate concepts, while term A and term D begin with a neutral superordinate concept, but the extensions of their definitions describe completely different characteristics of the concept in question.

6 Conclusion

The results of the preliminary analysis carried out for the purposes of this paper showed that 80 percent of multiple entries turned out to be synonyms and mainly belonged to the category of prototypical concepts that cannot be ascribed to the univocity principle. They therefore required additional terminographical work which included full cooperation among terminologists, Croatian standard language specialists and field experts. This type of data harmonization, beginning at term level, is time-consuming, rather demanding and potentially unsuccessful (in the case that field experts do not approve of the terminologists' proposals).

The results of the analysis suggest that, within the given framework of *Struna* termbase, cooperation among field experts and terminologists should be enhanced, as suggested by ISO 860.

Based on the experience gained so far, and considering both the organizational structure and workflow of *Struna*, future projects should roughly outline concept systems and preliminary glossaries at the outset and present them to their assigned terminologists before entering the intended number of terms into the termbase. This would allow harmonization to begin at concept level and continue at term level, thereby considerably easing the harmonization process and improving the quality of the termbase.

7 References:

1. Mihaljević, M., Nahod, B. Croatian Terminology in a Time of Globalization. In: Ledinek, N., Žagar Karer, M., Humar, M. (eds.) *Terminologija in sodobna terminografija*, pp. 17–26. Založba ZRC, ZRC SAZU, Ljubljana (2009)
2. Cabré, M. T. Terminology: theory, methods and applications. John Benjamins, Amsterdam; Philadelphia (2000)
3. Cabré, M. T. Elements for a theory of terminology: towards an alternative paradigm. *Terminology*: *international journal of theoretical and applied issues in specialized communication*, pp. 35–57, 6,1(2000)
4. ISO 860:2007 Terminology work – Harmonization of concepts and terms. International Organization for standardization, Geneva (2007)
5. UNESCO Guidelines for terminology policies: formulating and implementing terminology policy in language communities. UNESCO, Paris (2005)
6. Karsch, B. I. Why doublettes are bad. *BIK Terminology – solving the terminology puzzle, one posting at a time*. Available online at [http://bikterminology.com/2011/06/15/why-doublettes-are-bad/] (accessed March 26th, 2012)
7. ISO 704:2009 Terminology work – Principles and methods. International Organization for standardization, Geneva (2009)
8. Temmerman, R. Questioning the univocity ideal: the difference between sociocognitive terminology and traditional terminology. *Hermes*: *journal of linguistics*, pp. 51–91, 18(1997)
9. Karsch, B. I. Profile of a terminologist in localization environments. *The journal of internationalisation and localisation*, pp. 122–150, 1(2009)