The Croatian National Termbank STRUNA: A New Platform for Terminological Work

Maja Bratanić and Ana Ostroški Anić

Institute of Croatian Language and Linguistics, Department of General, Comparative and Computational Linguistics, Zagreb, Croatia

ABSTRACT

The development of the Croatian Special Field Terminology program (known by its Croatian acronym Struna) began in 2007 as part of an initial coordination project launched at the initiative of the Croatian Standard Language Council, and has since been financed by the Croatian Science Foundation. It is being carried out at the Institute of Croatian Language and Linguistics, which serves as the national coordinator. This paper describes the current design of the e-Struna termbank and explains the adjustments made in the database structure and in the terminographic approach, both to support and reflect the methodological issues concerning interdisciplinary and multidisciplinary work. Based on examples taken from the Croatian anthropological terminology collection special attention is given to two frequently neglected categories of terminological description: context and note.

Key words: terminology, terminological database, harmonization, context, note

Introduction

In 2007, as the first step in Croatian terminology planning, the Croatian Standard Language Council initiated the project «Development of Croatian Special Field Terminology» (referred to here by its Croatian acronym, Struna). The program has been financed by the Croatian Science Foundation and is being carried out at the Institute of Croatian Language and Linguistics, which was chosen to serve as the national coordinator. The termbank was officially inaugurated on the web in February 2012.

The aim of Struna is to gradually make standardized Croatian terminology for all professional domains available to the public by means of coordinating the work of domain experts, on the one hand, and terminologists and language experts, on the other. The broader objective of the program is to establish a framework for a national terminology policy and lay the foundations for more structured education in this field. The electronic database (e-Struna) developed for that purpose is a typical national terminology database intended as the primary tool for the implementation of a more or less explicitly recognized approach to terminology planning.

As in most normative terminological databases, the terminological work carried out in e-Struna is largely organized according to the principles of the General Theory of Terminology (GTT), which was inaugurated by Eugen Wüster, and codified in somewhat modified form in a number of ISO/TC 37 terminological standards. A firm methodological framework for practical terminographic work is necessary primarily because of the specific nature of the cooperation between the field specialists at one end of the process and terminologists and other linguistic experts at the other. A uniform approach to the terminological description guaranteed by a national terminology project environment is crucial to making the various subject fields in the database as structured and uniform in description as they possibly can be. On the other hand, the variety of the domain knowledge included in Struna, and the various characteristics of each domain – its conceptual structure and dynamics, specific communicative settings, and intended users – have called for the modification and adjustment of the general terminological principles in the terminographic description of particular domains.

According to the General Theory of Terminology and the traditional approaches in general, terminology is primarily a prescriptive activity with terminology standardization as its main goal geared towards unambiguous
specialized communication. The principal objective of terminology work is thus to eliminate ambiguity from special purpose languages and to establish terminology as an autonomous discipline. The basic tenets of terminology as envisaged by Wüster and his followers can be summarized in the following principles: terminology studies concepts before terms, i.e. its perspective is onomasiological, which distinguishes it from lexicography; concepts are clear-cut and have a fixed place in a concept system; a term is assigned permanently to a concept and it can be adequately defined by a traditional definition. In other words, each concept corresponds to only one term and one term refers to one concept only.

This approach remained dominant for more than fifty years, and in practical terminography it still prevails, but it often turns out to be too limited. Terminographic work needs a more explicit linguistic dimension and a broader perspective that must take into account the paradigmatic and syntagmatic structure of terminological units, the collocational behavior of terms and the syntactic patterns they follow, not only terms' synchronic but also their diachronic dimensions, their tendency to variation and, quite often, various sociocultural and extralinguistic information pertinent to their meaning and usage. Adhering closely to the principles of GTT built into ISO Standards – albeit in somewhat modified and less rigid form – can therefore sometimes pose problems even in a predominantly prescriptive approach like the one adopted by Struna.

The theories that arose in reaction to GTT in, roughly, the last two decades incorporate premises from sociolinguistics and cognitive linguistics and are by definition descriptive in nature and take into account the reality of term variation. In the attempt to analyze terms as they actually are used and behave in text these approaches focus on the social, communicative, and cognitive aspects of terminology.

Having adopted a well established organization of terminological information and by implementing the dominant standards for termbase design, Struna does not allow much deviation from the traditional approach, but at the same time makes full use of the means at hand.

In the first part of this paper (section 2), the underlying methodological approach and basic structure of the Struna termbase will be laid out, and the data categories that structure the conceptual and terminological description will be explained in more detail. The second part (sections 3 and 4) discusses certain challenges in the harmonization of entries across various domains and elaborates on how contextual information organized in the categories of context and note helps consolidate the conceptual information that the database contains.

The Terminological Database

Because Struna is by nature and organization a national terminology project, its termbase had to be envisaged as a normative termbase, a primarily monolingual terminology resource aimed at Croatian LSP terminology standardization. Nevertheless, even an essentially monolingual termbase can largely benefit from a multilingual (or at least bilingual) dimension, and there are several obvious reasons for this:

a) foreign language equivalents serve as an efficient control mechanism for the harmonization of terms across domains
b) this is a prerequisite of any data-sharing and termbase exchange
c) translators are very likely the most common termbase users.

Data categories

In the design of e-Struna, the main idea has been to make it as user friendly as possible, for both terminologists and final users. More importantly, since the project’s organization required a proper terminology management system, the termbase had to be planned to serve as both an editing and storage tool, and a search and retrieval tool. In order to make e-Struna compatible and exchangeable with existing termbases, its structure was designed in accordance with the TEI P5 guidelines for text markup and the TBX standard format for the representation and exchange of terminological data. The Struna termbase has been modified several times so far, primarily in terms of the number of data elements per record, as well as its functional properties. The current list of record elements consists of the following data categories.

1. subject field and subfield
2. preferred Croatian term (simple and complex – i.e., multiword expressions)
3. source of the term
4. foreign term language label
5. neologism label
6. interdisciplinary term label
7. grammatical information
8. definition
9. source of the definition
10. context
11. source of the context
12. synonyms – according to their normative status (admitted, proposed, deprecated, obsolete, colloquial)
13. equivalents (English, German, French, Italian, Latin, Russian, etc.)
14. subordinate concept
15. antonym
16. abbreviation (Croatian and other languages)
17. symbol
18. formula
19. equation
20. hyperlink
21. picture
22. note
23. correspondence (chiefly among the domain editors, terminologists, and language experts)
Apart from the above, there are four automatically generated data categories: term ID, creation date, created by, last modified by. The current stage of term processing is also indicated. These categories can be divided according to TEX data category standards into those on the termbase level, the concept level, the language level, and the level of the term. Data category specifications for ISO/TC 37, according to ISO 12620 from 2009, consist of two mandatory classes – the administration information section and the description section – with further enhancements made to the data category specification following the classes of the conceptual domain and the linguistic section.

Harmonization and Coordination

There is no generally accepted terminological description of concepts that could be used to equally represent the knowledge structure of all domains constituting a term base. (At this stage Struna encompasses twenty professional domains, some of which are still under construction or in the process of revision and thus are not yet open to public). Due to the rather unique model of termbase creation used for Struna, with the participation of both domain experts on one hand and terminology and language experts on the other, a certain level of discrepancy in terminological description has yet to be overcome.

Terminology work done on a large scale occasionally creates incongruities, both in term processing and terminology management. In the initial phases of Struna, insufficient coordination in delimiting various domains led to multiple records of equivalent terms in the termbase with identical (or roughly identical) meanings but different definitions, because of a number of terms that can justifiably be classified in more than one domain.

To avoid repetition and multiple records of the same concepts (defined in the same context), the category of interdisciplinary term was introduced, which enables the compilers to simultaneously attribute the term they consider relevant for their domain to its original domain (e.g. chemistry or physics).

It has become clear that harmonization within a particular domain and across domains must be an ongoing process, always defining and redefining when new concepts are introduced. Categories that are of particular use for harmonization across domains are the categories of context and note, because they contain extralinguistic and encyclopedic information about the concept, thus placing the concept more clearly in a communicative context, and making it easier for terminologists to compare detailed descriptions of related concepts.

Contextual Information in the Categories of Context and Note

All supplementary information relevant for a complete understanding of a defined concept is organized in e-Struna mostly into the categories of context and note. According to the General Theory of Terminology (GTT), the principles of which are largely applied in ISO standards, these categories «complement the definition but are not essential for differentiating the concept from other concepts»4. Therefore, they are not obligatory in a terminological entry, and can be left out when defining the concept. However, if one wants to define a term according to its conceptual frame and delimit its characteristics in relation to terms denoting similar or related concepts, both contextual and encyclopedic information is needed in its terminological description in order to provide the full communicative setting for the term and the concept it designates. This has proved to be particularly true in the processing of anthropological terminology, which often requires a deeper analysis of culturally embedded meanings, acknowledging a wide variety of theoretical interpretations. Many compilers have thus found these categories particularly useful for producing better descriptions of a certain concepts.

The category of context has traditionally been used in terminological descriptions as a source of relevant linguistic information, mainly to place the term in an appropriate syntactic or discursive setting. ISO standard 704, «Terminology work», defines context as every text that contains the term, but considers the defining context, «a context that allows the user to deduce the meaning of the concept by implication», to be the type of the context that should be used in terminology description. This approach is somewhat in line with the shift from using contexts mainly for linguistic description to making context a category that serves as a source of domain knowledge information on the concept it describes.

In the past twenty years traditional knowledge-poor contexts – that is, contexts that do not contain domain knowledge relevant to the term – have been replaced in terminological databases with knowledge-rich contexts, and in many bases also with contexts including knowledge patterns. Knowledge-rich context was first defined by Ingrid Meyer as a context «indicating at least one item of domain knowledge that could be useful for conceptual analysis»5. In the domain of anthropological terminology, the categories of context and note have been used extensively as a source of both relevant linguistic and extra-linguistic information. Apart from showing the use of the Croatian term in its linguistically relevant context (i.e., in a sentence), context also offers other concept-related information. Moreover, many entries show that the contextual information provided in this category makes up for a type of defining context, that is, a context that includes «all or most items of domain knowledge necessary to understand a concept»6.

According to ISO 1087-1, a note is very broadly understood as «a statement which provides further information on any part of a terminological entry»7. ISO 704 has a more precise description, stating that notes «may include non-delimiting characteristics or optional parts often associated with the concept, or typical elements that make up the extension of the concept which complement
the definition but are not essential for differentiating the concept from other concepts\(^4\). Optional characteristics associated with the concept are in many cases exactly those concept characteristics that do not define the concept, but place it in a broader context of the domain it is part of. Therefore, it is rather safe to assume that the categories of context and note contain in many terminological descriptions rather similar information, the types of which all fall under the umbrella of what is called contextual information.

Since the type and nature of the information given in the categories of context and note has not yet been delimited for the users working in Struna, the information that can be found in these categories greatly varies from one domain to another. Strictly speaking, the context is always taken from an actual written source, and the bibliographic information on that source is usually provided as well. The note, on the other hand, can be extracted from a source, but it is more often written by the field specialist himself or by the terminographer working on the same domain. Therefore, no place for recording the source of the note has been provided in the database structure.

The information that can be found in these two categories generally falls under three types: concept level information, term level information, and encyclopedic information. Several examples of these types of contextual information are presented below, based on terminological entries from the subdomain of linguistic anthropology. These entries are included in the analysis because they all contain information either in the context or the note, or in many cases in both categories. Since linguistic anthropology is an interdisciplinary field, this makes the description of its domain knowledge all the more demanding. The actual entries in e-Struna contain definitions, contexts and notes in Croatian only, but both Croatian and English versions of the terms and other parts of terminological description (relevant for a particular example) are listed here.\(^\ast\)

**Concept level information**

Concept level information extracted from the categories of context and note includes the following: (a) conceptual relations between the defined concept and other concepts in its framework, especially regarding the concept’s relationship with a superordinate concept which is not explicitly shown in the terminological description; (b) conceptual relations between the concepts in a subdomain or in an entire domain, including interdisciplinary relations; and (c) information on the multidimensionality of the concept.

The first proposed subtype of concept level information can be extracted from the terminological entry for the concept «discreteness», in which the information given in the note places «discreteness» in a conceptual framework among other features of the human language:

- (hrv) diskretnost
- (eng) discreteness
- (napomena) Charles Hockett uvrstio je diskretnost medju petnaestak temeljnih, odredbenih i općih obilježja ljudskoga jezika.
- (note) Charles Hockett included discreteness among fifteen basic, defining, and general features of the human language.

On the other hand, in the terminological entry «communicative event» the note is used to contextualize communicative event as a subordinate concept of «communicative behavior», along with «communicative situation» and «communicative act».

- (hrv) komunikacijski događaj
- (eng) communicative event / speech event
- (napomena) (...) tri su jedinice komunikacijskoga po- nanošanja pogodne za analizu: komunikacijska situacija, komunikacijski događaj i komunikacijski čin.
- (note) (...) three units of communicative behavior are suitable for analysis: the communicative situation, the communicative event, and the communicative act.

A definition of communicative event («a unit of communicative behaviour that takes place in accordance with social rules or norms for the use of speech») defines the concept as a subordinate concept of communicative behavior, but says nothing about its relations with other concepts in this conceptual framework.

The multidimensional nature of a concept is almost never explicitly shown in the definition, as in the example of the definition of «Broca’s area»:  

- (hrv) Brokino područje
- (eng) Broca’s area
- (definicija) dio lijevoga čeono mozdanog režnja koji sudjeluje u proizvodnji govora
- (definition) the part of left frontal lobe concerned with the production of speech

However, the semantic relations that refer to its multidimensionality can be extracted from the categories of context or the note, which describe both the concept and term evolution.

- (hrv) Brokino područje
- (eng) Broca’s area
- (kontekst) Pierre Paul Broca (1824. – 1880.) neurolog, tvrđio je 1861. g. da je afazija kod određenog pacijenta (čiji je mozak pregledao post-mortem) bila rezultat jedne točno omeđene lezije zadnje trećine drugog i trećeg čeo-

\(^\ast\) All translations were done by the authors. In certain examples, only those parts of the contexts or notes relevant for this discussion have been translated into English.
nog režnja (Head 1926:25). To posebno područje kasnije je označeno kao Brokino područje.

(cotext) Pierre Paul Broca (1824–1880), neurologist, claimed in 1861 that an aphasia in a certain patient (whose brain he examined post mortem) had been the result of an accurately circumscribed lesion of the posterior third of the second and third frontal convolutions (Head 1926: 25). This specific area was later named Broca’s area.

(napomena) Novija istraživanja proučavaju ulogu ovoga moždanog područja u širem smislu, kao sučelje između akcije i percepcije čija je središnja uloga u usklađivanju percepcijskih i motoričkih funkcija na kojima se temelji verbalna i neverbalna komunikacija.

(note) Recent research deals with the broader role of this region of the brain as an interface between action and perception, the central role of which is the coordination of perception and motor functions that serve as the basis of verbal and non-verbal communication.

The multidimensionality of the concept index is also not elaborated by the definition alone:

(hrv) indeks
(eng) index

(definicija) znak koji upućuje na označeno na temelju njihove prostore, vremenske ili uzročne veze

(definición) a sign that refers to the signified based on their spatial, temporal or causal relations

The category of context reveals that index can be understood from several perspectives – as a symptom, a trace of something or even as a result of an event – depending on the perspective of a person interpreting the index. Each of these perspectives reveals another dimension of the concept, one that is activated in a pre-set context or under particular conditions.

(hrv) indeks
(eng) index

(kontekst) (...) indeks može biti shvaćen kao posljedična kakva uzroka (npr. u slučaju simptoma kakve bolesti), kao najava neke mogućnosti (npr. oblacu su naznaka mogućnosti kiše), kao trag nekog počinitelja, kao odjek nekakva događaja i sl.

(context) (...) index can be understood as a consequence of a certain cause (e.g., in the case of a symptom of an illness), as a sign of some possibility (e.g., clouds are a sign of the possibility of rain), as a trace left by a perpetrator, as a public response to an event, etc.

**Term level information**

In the concepts belonging to the domain of linguistic anthropology this type of contextual information, in terminological entries in which it can be found, refers mainly to the use of the preferred term or its synonyms. The analyzed terms are placed in actual communicative settings (whether explicitly, as in the note relating to «linguistic landscape», or implicitly, as in the note for «ethnosemantics»), which adds communicative value to the terminological description.

(hrv) jezični krajolik
(eng) linguistic landscape

(napomena) Neki autori ovaj naziv upotrebljavaju za ukupnost uporabe različitih jezika u određenoj zemlji ili na nekome geografskom području.

(note) Certain authors use this term to represent the totality of the use of different languages in a certain country or in a certain geographical area.

(hrv) etnosemantika
(eng) ethnosemantics

(napomena) Nazivi etnosemantika, etnoznanost ili nova etnografijska sinonimi su za pristup proučavanju kulture kao znanja koji se razvio šezdesetih godina 20. stoljeća kao odgovor na dotad prevlađavajući materijalistički pristup.

(note) The terms ethnosemantics, ethnoscience, or new ethnography are synonyms for an approach to the study of culture as a body of knowledge that developed in the 1960s (...).

It is especially useful for the end user to learn that a certain term can be used within the same subdomain for a slightly different conceptual reality. This also means that this kind of information provides another perspective not only on the use of the term, but also on the multidimensionality of the concept itself. A definition, as traditionally understood, would omit this type of information as concept’s non-delimiting or non-defining characteristics, as can be seen in the definition of ethnosemantics:

(hrv) etnosemantika
(eng) ethnosemantics

(definicija) znanstveno proučavanje načina na koji ljudi nazivaju i razvrstavaju pojave u prirodnoj i društvenoj okolini

(definition) the scientific study of the ways in which people label and classify the social, cultural, and environmental phenomena of their world

**Encyclopedic information**

Information that is of encyclopedic nature is specific to the note category, which is primarily intended for various linguistic and extralinguistic information about the terms,7 such as the use of a Croatian term in a certain sub-domain, or the use of spelling variants in English equivalents. However, it is also the category where non-delimiting characteristics or «typical elements that make up the extension of the concept»8 can be entered to better illustrate the concept. Therefore, encyclopedic information refers both to the term and to the concept level, although in the domain of anthropology within Struna, it mainly refers to a further illustration of the
concept. Term related encyclopedic information concerns the term’s creation, as is the case in the notes written for the terms Broca’s area and erasure:

(hrv) Brokino područje
(eng) Broca’s area

(napomena) Područje je dobilo naziv prema Pierreu Paulu Broci, francuskom lječniku i antropologu koji je u 19. st. otkrio da oštećenje toga dijela mozga uzrokuje probleme u gramatičkoj obradbi tijekom proizvodnje govora, dok za jezično razumijevanje uglavnom nema negativnih posljedica.

(note) This area was named after Pierre Paul Broca, the French doctor and anthropologist who discovered in the 19th century that (…).

(hrv) brisanje
(eng) erasure

(napomena) Naziv su osmisile J. Irvine i S. Gal za semiotički proces ikonizacije koji pojednostavljuje jezičnu i društvenu stvarnost izostavljanjem činjenica i pojava koje se ne uklapaju u ideološki okvir.

(note) This term was coined by J. Irvine and S. Gal to designate the semiotic process of iconization that simplifies language and social reality by omitting facts and activities that do not fit into an ideological framework.

Encyclopedic information that refers only to the concept can be a more specific definition of the concept – i.e., a flexible definition that restricts the meaning of a concept to a particular contextual domain. An additional definition is often supported by a sentence that states the field of application of the concept, especially if it is a method or a process. The difference between a definition and an additional definition in the form of a note can be seen in the terminological entries « ethnopoetics » and « language policy »:

(hrv) etnopoetika
(eng) ethnopoetics

(definition) analysis and interpretation of formal organization of texts and their intertextual relations in oral literature, special regard to the context and performance

(napomena) Etnopoetika podrazumijeva jezični opis s posebnim osvrtom na poeticske strukture u govorima koje mogu upozoriti na dodatne slojeve značenja, a koje se ne mogu spoznati drugim metodama. (…) Etnopoetika se primjenjuje u proučavanju poezije, mitova, narodnih priča i naracija.

(note) Ethnopoetics implies a linguistic description with a particular reference to those poetic structures in speech that can show additional layers of meaning (…) Ethnopoetics is applied in the study of poetry, myths, folk tales, and narratives.

Further information found in notes often contains exemplification of information contained in the definition:

(hrv) komunikacijska situacija
(eng) communicative situation

(definition) a sequence of social activities in a specific time or space or an instance of communication

(napomena) Primjeri su komunikacijskih situacija vjerski obred, sudski proces, zabava, aukcija i sl.

(note) (…) Examples of communicative situations are a religious ceremony, a court process, a party, an auction, etc.

Notes can also contain any information on a change or a particular process that a concept undergoes, the result of which is a change in the concept’s characteristics, influencing the way in which we perceive its scope of meaning today.

Concluding Remarks

There are a number of reasons why doing terminology work on a national termbase level may seem advantageous over terminology management in the context of specialized professional or domain termbases. The collaboration between domain experts and terminologists assures a high quality of terminology work in terms of faster term validation, faster dissemination of new terms, and easier access to various sources of information relevant for terminology documentation. A single common termbase also makes for easier and quicker access to terminology from a larger number of domains.

On the other hand, concept overlapping is unavoidable – making the harmonization and standardization of terminology more difficult – but such overlapping nevertheless occurs in termbases developed for a single domain as well. Different domains have different needs, both in terms of their specialized discourse and regarding the nature of their domain conceptualization, which
makes concept harmonization challenging and in a few cases impossible.

Another drawback is that having a large number of people working in the termbase makes the coordination between projects increasingly more difficult, but as Nilsson points out, terminology work in a national termbank is therefore more efficient because it is easier to re-use existing terms and definitions.

In order to further raise the level of terminological quality, term records have to be checked and updated whenever a related concept is introduced or the meanings of existing ones change with time. Terms and definitions can be updated more easily by automated procedures, and thus a dynamic termbase is the next level that should be aspired to.

In its more advanced phases of establishing standardized terminology for various professional domains, it is hoped that e-Struna will gradually improve the circulation of knowledge and information in the Croatian language, as well as in the broader multilingual environment. We see this as a necessary prerequisite for the very existence of Croatian as a standard language, particularly with regard to its becoming one of the official languages of the European Union.

Although originally designed and built as a highly normative terminological database, e-Struna has stepped outside its initial boundaries, and today it contains a variety of information in its terminological entries, providing a good descriptive approach and taking the first step towards the creation of a knowledge base.

Acknowledgements

This research has been done within the project Struna – Development of Croatian Special Field Terminology, which is supported by the Croatian Science Foundation and carried out at the Institute of Croatian Language and Linguistics.

REFERENCES


M. Bratanić

Institut za hrvatski jezik i jezikoslovlje, Republike Austrije 16, 10 000 Zagreb, Croatia
e-mail: mbratan@ihjj.hr

STRUNA: HRVATSKA NACIONALNA TERMINOLOŠKA BAZA – NOVO POLAZIŠTE ZA TERMINOLOŠKI RAD

SAŽETAK

Program Izgradnja hrvatskoga strukovnoga nazivlja (Struna) započeo je 2007. godine na inicijativu Vijeća za normu hrvatskoga standardnog jezika, a Hrvatska zaklada za znanost financijski ga podupire od 2008. godine. Program se odvija u Institutu za hrvatski jezik i jezikoslovlje koji je izabran za nacionalnoga koordinatora. Cilj mu je osigurati pretpostavke za normiranje hrvatskoga nazivlja što većega broja stručnih i znanstvenih područja i disciplina. U radu se opisuje struktura terminološke baze STRUNE, objašnjava terminografski pristup i ilustriraju metodološki problemi koji se javljaju u interdisciplinarnome i višedisciplinarnome okružju. Na primjerima naziva obrađenih u STRUNI u okviru projekta Izgradnja temeljnoga nazivlja u antropologiji pobliže se razmatraju dvije često zanemarene kategorije terminološkoga opisa – kontekst i napomena.