РУССКО-АНГЛИЙСКИЙ ТЕЗАУРУС ПО КОМПЬЮТЕРНОЙ ЛИНГВИСТИКЕ

Соколова Е. Г. (minegot@rambler.ru),
Российский государственный гуманитарный университет,
Москва, Россия

Кононенко И. С. (irina_k@cn.ru),
Институт систем информатики имени А. П. Ершова СО РАН,
Новосибирск, Россия

В статье обобщается опыт двухлетней работы по созданию русско-английского информационно-поискового тезауруса по компьютерной лингвистике (КЛ). Дается обоснование ввода в описание терминов ссылок на разделы КЛ и смежных наук. Коротко описываются типы информации в словарной статье термина. Обсуждаются основные проблемы описания терминов, связанные с такими особенностями тезауруса, как двуязычность и незрелость представляемой им области знаний. Возникшие терминологические проблемы анализируются с использованием системы классификационных признаков.

Ключевые слова: тезаурус, компьютерная лингвистика, русско-английский, определение термина, дескриптор, отношение.

RUSSIAN-ENGLISH THESAURUS ON COMPUTATIONAL LINGUISTICS

Sokolova E. G. (minegot@rambler.ru)
Russian State University for the Humanities, Moscow, Russia

Kononenko I. S. (irina_k@cn.ru)
A. P. Ershov Institute of Informatics Systems SB RAS,
Novosibirsk, Russia

This paper summarizes the experience in the construction of Russian-English information retrieval thesaurus on Computational Linguistics (CL). The need for relating thesaurus terms to the subareas of CL and adjacent sciences is substantiated and the hierarchical structure of subareas is discussed. The kinds of information given in the thesaurus term entry are outlined. A number of terminology description issues are discussed with regard to the specific features of the constructed thesaurus such as bilinguality and insufficient development of Russian CL. Terminological problems are analysed using classification parameters.

Key words: thesaurus, computational linguistics, russian-english, term definition, descriptor, relationship

Introduction

We discuss the results of a two-year (2010–2011) project[1] to develop bilingual Russian-English thesaurus on Computational Linguistics (CL). The project presents high interest in practical (lexicographic) and theoretical aspects taking into consideration the absence of more or less representative descriptions of Russian CL terminology. The closest to the topic works available to date are thesaurus by Nikitina (Nikitina, 1978) that is small and out of date and recent thesaurus INION (Smirenskii, 2007). They describe a small number of CL terms and a lot of purely linguistic terms which have no relation to information and text processing. Other limitations of both thesauri are the lack of definitions of terms and monolinguality. Russian term definitions are given, for example, in the Glossary on Artificial Intelligence (AI) (Averkin, 1992), this field being considered as subordinating or intersecting with CL, but common terms often differ in semantics, for example, *syntactic analysis* in CL and in AI. CL, and especially Russian CL, is not mostly methodological science, as opposed to AI[2].

[1] The research was supported by Russian Foundation for Humanities, project No 10-04-12108в.

[2] This also explains the fact that we include subarea of theoretical CL — Generative Grammar — which is related to syntactic analysis task, syntactic annotation task, the task of modelling word order in NLG, etc., and don't include "formal semantics" which corresponds rather to methodological means. See also sec. 2 below.

CL as a research area came into being after the appearance of computers and immediately began reconsidering current descriptions of natural languages (NL) for the purpose of creation of machine translation (MT) systems. Since then CL has extended to other applied areas — Information Retrieval (IR), Information Extraction (IE), etc., and now it remains a new research field whose changes accompany the development of technical devices (computers, global network), scientific progress (linguistics, mathematics) and social advance that affects information technologies. CL is an open area whose theoretical field intersects with linguistics and psychology.

Bilinguality in our thesaurus is symmetric in the sense that every Russian term is provided with English equivalent and vice versa. But the meanings of Russian and English terms often diverge or no equivalence is at hand so that we have to resort to translation. Really, Russian and English states of the art in CL are different. The methodological and technological lagging of Russian CL with regard to the world level is observed on the site of the annual "Dialogue" conference which is the only one in Russia to represent the field of CL in more or less full extent. Also, in Russian there are no textbooks or manuals on CL that would present the area in full enough details, without them being too subjective or compiled from English sources.

Our primary task was to represent the **original Russian CL terminology** and map it into the English one. The goal was not to "translate" Russian CL terminology, but to show it and then merge it with the worldwide state-of-the-art CL. Accordingly, for the Russian part of the thesaurus the collection of proceedings presented at the "Dialogue" Conference in 2000–2010[3] was created and analyzed, it being a helpful source of Russian terms in real use. Proceedings of the International conference "Corpus linguistics" and the manual on corpus linguistics (Zakharov, 2005) served as another source of Russian terms in this one of the most significant subareas of CL. But some empirical and technologically advanced subfields such as Speech Technologies were analyzed mainly in the reverse direction because a lot of terms are lacking in Russian, so the English-language terminological sources have been used and their terms and definitions translated into Russian. As for the English part, dictionaries as well as indices and glossaries of textbooks and manuals have been looked for English terms and definitions.

In (Sokolova et al., 2011) we described the initial phase of the development of Russian-English thesaurus on CL and discussed such questions as the choice of candidate sources of terms, the techniques of machine-aided terminology extraction and selection of basic term list.

In this paper we present the result of the project — bilingual Russian-English thesaurus on CL — section 1; in section 2 the scopes of CL subareas are analyzed in the dynamic perspective; section 3 outlines the types of information given in the thesaurus term entry; in section 4 terminological problems are observed in connection with specific features of the thesaurus and their analysis is given in terms of classification parameters.

[3] Proceedings of the International Conference "Dialogue" are available at: http://www.dialog-21.ru/

1. Results of the project

Results of the project are twofold: a) a list of terms and their descriptions that are usually given in thesauri to represent research area terminology, for example in the above AI thesaurus; b) structuring of CL into subfields and relating terms to subareas or adjacent areas. So the main results of the project are:

- **Russian-English Thesaurus on CL** (**REThes-CL**) exists and is available at: **http://uniserv.iis.nsk.su/thes/**; its content — the list of terms, their definitions, and relations — is open to discussion;
- the content presenting 1671 terms (1031descriptors) on different directions of CL is representative enough to be the base for the thesaurus technology;
- description of CL area organization in the context of adjacent sciences that is not usual in standard thesauri;
- instruction for the term description presenting the specific technology of REThes-CL;
- experience drawn while constructing the thesaurus on CL, which can be characterized as a "scientific — practical area"[4] rather than a "science".

In the next sections we consider the last three points in more details.

2. CL area, its subareas and adjacent sciences

At the initial stage of the project an important problem was missed that is rooted in the abundance of research directions and interdisciplinary nature of CL.

We see that definitions of CL are very vague and in a sense "negative" since they describe CL as "something existing BETWEEN or INTERSECTION of Linguistics, Computer Science and AI (see the term *computational linguistics* in REThes-CL). They don't show the CL true object — processing information presented via NL. Text and speech forms of NL communication plus diversity of directions of text processing divide the CL area into subareas. On the other hand, we can't ignore the adjacent sciences which intersect with the CL proper and form the context for the entire area of CL. The resultant hierarchy of CL subareas and adjacent sciences together with terms for them (top terms) are presented on the website of REThes-CL.

The extensiveness of CL and diversity of its subareas lead to incomprehensibility of certain terms for the users. The parenthetical qualifier is sometimes used to give reference to a subarea, e. g. *pattern (*in *information extraction)*. But the qualifier's main function is to remove the ambiguity of terms, so it should be used only in case of homonyms. The problem of homonymy existing between terms from different subareas of CL (or from CL and adjacent areas) is resolved by introduction of explicit relationship between a term and subarea(s). This ensures for a fully coherent picture of the subarea terminology, which becomes easily accessible for the interested user, on the

[4] Term of O. F. Krivnova

one hand, and could be useful in the indexing purposes to allow for a more precise/accurate search, on the other.

Multiple relationship to subarea or adjacent science is also important due to the fact that subareas are changing over time, so new relations of a term can be added. There are two opposite tendencies of change: generalization — when a separate science forms, and derivation — for child subarea. Both tendencies are observed in IR subarea of CL, historically second after MT. IR is specified in the following core definition: "as an academic field of study, IR … is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections" (Manning et al., 2008). Now IR acquires features of separate science: there appear manuals, courses in the universities, terminological dictionaries, e. g. "SE/SEO glossary"[5].

On the other hand, the traditional subtasks of IR involve structuring and rearrangement of selective information, be it explicit or implicit in the documents: classification/clustering, summarization, IE. In the course of time, however, the development of techniques and tools has led to the establishment of separate applications in this area, independent on the IR task proper. This illustrates the second tendency which leads to coexistence of broad and narrow interpretations of these subareas and their top terms. Thus, in (Grishman, 1997) two definitions of IE subarea are given: 1) a broad view of IE: any method of filtering information from large volumes of text. This includes the retrieval of documents from collections and the tagging of particular terms in text; 2) a narrower definition: the identification of instances of a particular class of events or relationships in a NL text, and the extraction of the relevant arguments of the event or relationship. But in the modern documentation for Gate (Cunningham et al., 2011) this subarea is interpreted as strictly separate from IR: IE differs from IR and traditional techniques in that it does not recover from a collection a subset of documents. Instead, the goal is to extract from the documents salient facts about prespecified types of events, entities or relationships.

An interesting result of the progress over time is the division of CL, which used to be a purely applied science, into Theoretical CL and Applied CL that is made by Hans Uszkoreit (Uszkoreit, 2000). This division is absolutely necessary to differentiate between Applications and models imitating elements of processing speech/texts in human mind (functions of "black box" in cybernetics). Division into subareas in Theoretical CL follows the principle used in Applied CL — we define subarea when there are technologies to resolve **a task**, not following any level-based "theories" of NL in CL, which are very different. We declare the subarea of Syntactic Analysis because there are technologies for the task including dependency filtering method, unification based syntactic-semantic approaches, and others, but there are only experimental attempts in Semantic Analysis. In spite of many experimental or theoretical attempts in this area, especially in the 70–90-es, including R. Shank and many others, to-day the only type of "purely" semantic representations available in the Internet (at least in Russian CL) are "graph structures" in terms of semantic analysis of N. N. Leont'eva (Leont'eva, 2006) available at www.aot.ru. And this approach is close to the tasks of IE.

[5] available at: http://www.infonew.ru/seo_glossary.php#39

3. Representation of terms in the thesaurus entry

The main information items of REThes-CL are domain terms. They are represented by two types of lexical units: single words (mainly nouns) and nominal phrases. The variety of terms is broken down into descriptors (preferred terms) and non-descriptors (variant terms that include synonyms, lexical variants, quasi-synonyms, and abbreviations). The descriptor is chosen as a representative of equivalence class of terms that refer to the same concept and thus exhibit an equivalence relationship.

Terms of any type are provided with ***term name***, ***language***, ***author***, and ***comments***.

The relevance of any term is evidenced by relating it to terminological ***source*** (documents or text collections), this being an additional authority information concerning terms. Where terms or definitions have been extracted from specific available references (glossary, subject matter index, or text), the reference name(s) are given within the term entry. For the collection type sources the ***frequency of occurrence*** is specified.

For the descriptor terms additional attributes are as follows: ***term definition***, ***subarea***, and ***qualifier*** (which is a part of descriptor). The qualifier may refer the term meaning to conceptual category or subject domain, e. g. *accent (pronunciation)* and *accent (prosody), token (corpus linguistics)* and *token (informatics), разметка текста(процесс)*= 'text tagging' and *разметка текста (объект)*= 'tagged text'. Term definitions are not standard for information retrieval thesauri, but it makes the thesaurus a good source of CL knowledge. Definitions are mostly drawn from existing glossaries, papers and manuals, and references to the sources are given. Sometimes one term can have two or three definitions taken from different sources. The list of major sources for terms and definitions is presented in (Sokolova et al., 2011).

The REThes-CL entry presents information concerning interrelations between terms, with following basic types: the above mentioned ***equivalence***, ***hierarchical***, and ***associative***.

Specialized equivalence relationships cover different types of correspondence between descriptors and non-descriptors. ***Synonymy*** holds if a descriptor is substituted for a non-descriptor in all contexts: e. g. *valency* replaces *valence* and *subcategorization; semantic role* substitutes for *semantic case, thematic role, deep case, case role, theta role,* etc. ***Alternative synonymy*** or ***combination synonymy*** are used to replace a non-descriptor by multiple preferred terms with a relation OR or AND in between. For example, *topic* (='a particular subject that the text discusses') and *patient* (='the semantic role of an entity that is not the agent but is directly involved in or affected by the happening') are alternatively used to replace the ambiguous non-descriptor term *theme*. The non-descriptor term *statistical machine translation system* is replaced by a combination of descriptors *machine translation system* and *statistical machine translation*.

Hierarchical and associative relationships hold between the descriptors of each monolingual part of REThes-CL. The ***Hierarchical*** relationship is a connection between broader and narrower terms in the thesaurus. ***Broader*** and ***Narrower***

relationships are further differentiated using three subtypes: ***Generic***, ***Instance***, and ***Partitive***.

A set of descriptors subordinated to the same immediate broader term may be grouped under a ***Hierarchy note*** which specifies a characteristic of hierarchical division of the broader term: *machine translation* is subdivided into *example-based machine translation*, *rule-based machine translation*, and *statistical machine translation* by 'approach' and into *fully automatic translation* and *machine aided translation* — by 'degree of human involvement'.

Associative relationship denotes any nonhierarchical semantic relationship that holds between the descriptors referring to closely related concepts (action/product of action, cause/result, concept/property, agent/counter-agent, etc.).

Translation equivalence relationship connects the descriptors in different languages that refer to the same concepts.

4. Terminological problems arising from immaturity of the Russian CL and bilinguality of thesaurus

Our work to construct REThes-CL has highlighted the effects of immaturity of Russian CL and reflected the unbalanced development of Russian and English CLs, which is the heritage of separateness in the past.

The number of non-preferred terms (synonyms) is small in thesauri for established terminology systems. Our experience with REThes-CL shows another picture: the balance between descriptors and non-descriptors is roughly 2/1 (total descriptors — 1031 and total non-descriptors — 640). As a result of lagging of the Russian CL as compared to the English CL the 'no equivalence' situation, i. e. the absence of Russian terms, frequently occurs.

To analyze and solve the problems with terms in the process of bilingual thesaurus construction the consideration of certain relevant features seems to be helpful. They are enumerated below in relation to the Russian CL:

A. Novelty of a concept — {"+A": new concept, "–A": existing concept};
B. Existence of one, several or zero term name(s) for a concept — {"+B": one term, "++B": several terms, "–B": no term = a new concept borrowed};
C. Existence of stable relation between a concept and a term name (in case of a new concept the relation may be unstable, that results in coexistence of several term variants none of which is preferred by the scientific community) — {"+C": stable term, "–C": unstable term};
D. Potential for the hierarchical semantic change (shift) in the meaning of a term (change from the lower level of the generic or partitive hierarchy to the immediate higher level, or vice versa) — {"+D": possible shift, "–D": impossible shift}.

These parameters are mutually dependent and not all the combinations of their values are real: e. g. <+B, –C> and <– A, – B > are impossible. Other important combinations are presented below.

<(+A|-A), ++B, +C, (+D|-D)> — the descriptor choice problem. The choice is made on base of statistics: traditionally accepted term is selected for indexing purposes, e. g. *актант*, whose frequency of occurrence[6] is 733, is preferred to *аргумент* (616); *валентная структура* (20) is preferred to variants *валентная рамка* (14), *валентностная структура* (3), *схема валентностей* (3).

<+A, ++B, –C, (+D|-D)> — the term for a new concept choice problem. Besides statistics, the problem solving additionally requires the linguist/expert knowledge or intuition, e. g. for the *translation memory* some experts would prefer *архив переводов* (1) in spite of the higher frequency of *переводческая память* (8) and broad use of the calque *память переводов* (0) by translators-practitioners.

<+A, –B, –C, (+D|-D)> — 'no equivalence' in Russian CL and the necessity to coin the term, e. g. the term *целевой фрейм* found in (Kormalev, 2004) as a good match for *template* (used in the IE field to denote the final, tabular output format of IE) and *автоматический перевод устной речи* as a match for *spoken language machine translation*. Note that a lot of terms relating to Speech Technologies have been coined by the entry authors.

<(+A|-A), (+B|++B|-B), +C, +D> — the problem of semantic shift in cases of *модель управления* and *валентная структура* in their different (broad and narrow) meanings. Real use cases of the term *валентная структура* show the ambiguity of this term (quite similar to that of *модель управления)*, as they have both narrow interpretation (syntactic valencies of the predicate word) and broader interpretation (correspondence between semantic valencies and syntactic valencies). The problem is solved by introduction of the qualifier in the term name: *валентная структура* (without qualifier) and *валентная структура (синтаксис)*.

We consider the numeration of possible situations to be a convenient technique to monitor and solve terminological problems in the bilingual thesaurus construction.

Conclusion. Applicability of thesaurus

The thesaurus is a model of the research field in progress, not a description of fixed domain. The developers of REThes-CL face at least two interrelated problems: the bilingual character of the study and the immaturity of the research field. The thesaurus construction helps understand the current state of the art in Russian CL and speed up progress in this area.

REThes-CL is currently used in teaching CL to undergraduate students. It serves as a guide for the students to study the CL structure and state, to familiarize with terms actually in use in the field and to conduct real lexicographic research by describing terms.

Thesaurus will be effective for more accurate definition of basic notions of Russian CL and support mutual integration of CL studies in Russia and abroad; also it will be a good help to the researchers and translators of the field-related literature.

[6] Here the frequencies of occurrence in the Dialogue collection are given, and 'zero' means that the term variant has been found elsewhere but not in the Dialogue texts.

REThes-CL can be characterized as “monitor”, i. e. permanently enlarged under supervision. The opinion is relevant to it: “Like a taxonomy, a thesaurus is never "finished." New findings, and reinterpretation and restating of what is already known, require that terms be added, changed, and occasionally deleted. Continued usefulness of the thesaurus requires an ongoing commitment to updating.” (Milstead, 1998).

On the other hand, we hope that our thesaurus will be useful to solve library catalogs’ problems in describing CL, it being a new scientific-practical area of research. Information in REThes-CL is organized in accordance with standards for information retrieval thesauri, so it may serve as the authority base for indexing and content retrieval of CL texts.

References

1. *Averkin A. N., Gaaze-Rapoport M. G., Pospelov D. A.* (1992). Tolkovyi slovar' po iskusstvennomu intellektu [Glossary on Artificial Intelligence].Moscow, Radio and Communication, 256 p. Available at: (http://www.raai.org/library/tolk/aivoc.html)
2. *Cunningham H., Hainard D., Bontcheva K.* (2011). The Gate Uzer Guide: Text Processing with GATE (Version 6). University of Sheffield, Department of Computer Science. Available at: http://gate.ac.uk/
3. *Grishman R.* (1997). Information extraction: Techniques and challenges. In: Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, International Summer School, (SCIE-97), Springer-Verlag, pp. 10–27.
4. *Kormalev D. A.* (2004). Application of Machine learning to Text Analysis [Prilozheniia metodov mashinnogo obucheniia v zadachakh analiza teksta]. Trudy mezhdunarodnoj konferencii «Programmnye sistemy: teorija i prilozhenija» (Proceedings of International Conference “Programm Systems: Theory and Applications”), IPS RAN, Pereslavl'-Zalesskii, Moscow, Vol.2, pp. 35–48.
5. *Leont’eva N.N.* (2006). Avtomaticheskoe ponimanie teksta: sistemy, modeli, resursy [Automatic Text Understanding: Systems, Models, Resources]. Moscow, Akademiia, 304 p.
6. *Manning C. D., Raghavan P., and Schütze H.* (2008). Introduction to Information Retrieval, Cambridge University Press.
7. *Milstead J. L.* (1998). NISO Z39.19: Standard for Structure and Organization of Information Retrieval Thesauri. Paper presented at the Taxonomic Authority Files Workshop, Washington, DC. Available at: http://www.bayside-indexing.com/Milstead/z39.htm
8. *Nikitina S. E.* (1978). Tezaurus po teoreticheskoi i prikladnoi lingvistike. [Thesaurus on Theoretical and Applied Linguistics]. Moscow, Nauka.
9. *Smirenskii V. B.* (2007). Informatsionno-poiskovyi tezaurus INION po iazykoznaniiu. [Information retrieval Thesaurus on Linguistics by INION]. Moscow, INION, 200 p.

10. *Sokolova E. G., Semenova S. Yu., Zagorulko Yu.A., Kononenko I. S., Zakharov V. P., Krivnova O. F.* (2011). Selection and Preparation of Terms for Russian-English Thesaurus on Computational Linguistics. [Osobennosti podgotovki terminov dlia russko-angliiskogo tezaurusa po komp'iuternoi lingvistike] Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialogue 2011" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2011"). Moscow, pp. 644–655.
11. *Uszkoreit H.* (2000). What is Computational Linguistics? Available at: http://www.coli.uni-saarland.de/~hansu/what_is_cl.html
12. *Zakharov* (2005). Korpushaia lingvistika [Corpus Linguistics]. Saint Petersburg, 2005, 48 p.