***L.N. Beliaeva, O.N. Kamshilova***

**MT Results and Parallel Scientific Text Corpora for Lexicography**

**Abstract.** The paper suggests that the use of full-text parallel corpora for lexicographic and terminographic aims may turn more effective, provided they have a «built-in» corpus of MT results, since analysis and comparison of these corpora will make it possible to identify lexical units that are to be considered as dictionary entries. A corpus of MT results, we assume, can store the history of NP transformations and modifications in the corpus. The complicated structure of multicomponent terminological NPs, their variants and modifications within the same text determine the need for a three-part text corpus, including parallel/comparable texts and their MT translation.

**Keywords.** Lexicography, terminology, parallel and comparable corpora, MT, multicomponent NPs

**1.  Introduction: Applied Lexicography and Parallel Corpora**

The aim of the paper is to propose a way to optimize the use of parallel and comparable corpora as lexicographic resources by means of MT procedures and results. Being a major branch of applied linguistics, applied lexicography traditionally aims at building and updating subject oriented databases and automated/automatic dictionaries, specifically terminological ones. The level and reliability of information/knowledge extracted from texts of various composition, structure and destination is determined by the completeness and adequacy of lexicographic systems used for the purpose.

It has become almost commonplace, that much of lexicographic (terminological) job today is based on text corpora, that provide a reliable database for dealing not only with research issues, but with practical lexicographic tasks as well, such as terms identification and extraction, translation, etc. [Beliaeva 2009, 2014; Delpech, Daille 2010; Heja 2010; TTC Project[[1]](#footnote-1)]. Parallel and comparable text corpora are effectively used for creating multilingual lexicons and concordances.

In this paper we dare to suggest that, if we use full-text parallel or comparable corpora as a lexicographic base, it is necessary to expand them with a corpus of MT results. Analysis and comparison of these corpora will make it possible to identify lexical units as candidates for special dictionary entries [*cf.*: Delgado et al. 2002; David, Curran 2007; Lavie et al. 2008]. The main difficulty in this identification process is to establish the boundaries and structures of these lexical units. Whenever we deal with a scientific text, we have to admit that the lexical units in question are multiword terminological constructions. Scientific texts abandon in simple (without preposition) noun phrases (NPs) which are usually multiword units with a number of attributive elements modifying the head noun. They form one syntactic group with its head element and have a common syntactic function in the sentence.

NPs are objects of study in both theoretical and applied aspects [Baroni, Zamparelli 2010; Bergsma, Wang 2007]. Such phrases are functionally equivalent to a word, but at the same time they represent a convolution of a sentence, i.e., they are definitely units of syntax, not lexicon. The NP dependency structure has always been a major issue for MT or human translation of scientific texts (*cf.*: [Feldman, Dagan 1995; Babych, B., Hartley 2002; Shen et al. 2008; Reiter, Frank 2010]), especially when translating from English to any inflectional language. We assume that the internal structure of an NP correlates with the internal dependencies structure of the appropriate sentence. The point is to find a procedure to recognize this structure in a convolution.

The paper focuses on multicomponent terminological NPs in scientific texts, their structure and transformation to find a procedure for matchings in parallel and comparable texts.

**2. Tracking Noun Groups in a Scientific Text**

An NP in its simplest form consists of a determiner and a head noun or pronoun. It may consist of one word or include a number of embedded premodifiers that make it a complex unit. NPs with a number of premodifiers are called simple if they include no preposition, no matter how many premodifiers they have [Malakhovskaya et al. 2021]. As terminological NPs actually represent a sentence compression (convolution), their internal structure must correlate with the internal structure of a corresponding sentence, thereby revealing the syntactic dependencies. Thus, finding a procedure to recognize this structure in a concise form of an NP becomes a key problem.

Since an NP is a sentence convolution, the markers of relations between its actual components normally show in inflectional languages, while a simple English NP hardly has any, except for *-’s*. The compression of sentence structure, the external simplification of both structure and form of English NPs causes semantic complication.

Our corpus findings in parallel and comparable scientific text corpora of different subject areas (medicine, space systems, seismic isolation, power plants construction, machine translation, language teaching), built for research and practical translation and lexicographic aims, prove that most frequent in English are 2-component combinations with a head noun, which three times exceed three-element combinations, second frequent combinations in scientific and technical texts (see Table 1).

*Table 1.* **Frequency of English NP Length in Scientific Texts (Subject Area «Seismic Protection»[[2]](#footnote-2))**

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **NP length** | **Number of different models** | **NP frequency** |
| 1 | 2 | 1516 | 3457 |
| 2 | 3 | 674 | 1053 |
| 3 | 4 | 207 | 380 |
| 4 | 5 | 51 | 61 |
| 5 | 6 | 20 | 164 |
| 6 | 7 | 2 | 2 |
| 7 | 8 | 5 | 5 |
| **Total** |  | 2475 | 5122 |

In spite of their model variety, 2-component NPs prove to exceed other multicomponent NPs in scientific texts across different subject areas. Table 2 illustrates the distribution of 2-, 3- and 4-component NPs in a special comparable corpus on web and linguistic technologies:

*Table 2.* **Models of English NPs in Scientific Texts (Subject Area «Web and Linguistic Technologies»[[3]](#footnote-3))**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model number** | **Model** | **Length** | **Frequency** | **Number of NPs** |
| 1 | A+N | 2 | 1474 | 748 |
| 2 | A+PII+N | 3 | 9 | 6 |
| 3 | N1+N2 | 2 | 1407 | 530 |
| 4 | N1+N2+N3 | 3 | 248 | 128 |
| 5 | A/N1+N2 | 2 | 71 | 47 |
| 6 | N1+G/N2 | 2 | 24 | 18 |
| 7 | A1+A2/N1+N2 | 3 | 10 | 10 |
| 8 | A+G/N2 +N2 | 3 | 7 | 4 |
| 9 | A1+A2+N | 3 | 151 | 104 |
| 10 | A+N1+N2 | 3 | 292 | 172 |
| 11 | PII+A+N | 3 | 25 | 20 |
| 12 | PII+N | 2 | 170 | 73 |
| 13 | A1+N1+A2/N2+N3 | 4 | 3 | 2 |
| 14 | A1+C+A2+N | 4 | 15 | 9 |
| 15 | A1+A2/N1+N2+N3 | 4 | 6 | 7 |

However, the external simplicity of the most frequent English NP structures is misleading due to the fact that this simplicity could be the result of another NP or a sentence compression, which, as it was above mentioned, leads to its semantic complication.

Pursuant to this, NPs formation in a text is either merging NPs and single (independent) lexical units in a new, more complicated nominative construction, or condensing multicomponent NPs at the expense of deleting the units which are implicitly obvious.

In accordance with this we find two ways of building new NPs in a real text: either by adding a lexeme to a previously used or standard NP, thus producing a novel, more complicated nomination: *machine translation => machine translation method*, *machine translation service, machine translation program*, or by deleting implicitly obvious units, thus condensing the sentence structure to a multicomponent NP: *syntactic dependency, syntactic formalism, syntactic dependency tree annotation => dependency annotation formalism*

The first way is a step-by-step process of gradual transformation (complication), adding specific characteristics to the head element, while the second presents a transformation process of sequential convolution which is successively realized on three levels:

Level 1. Transfer from a complex NP (with a preposition) to a simple one by inversion of its elements: *phrase-structure trees from dependency annotations => syntactic phrase-structure dependency trees annotation.*

Level 2. Elimination of duplicate components in the new NP: *syntactic phrase-structure dependency trees annotation* =*> syntactic dependency tree annotation.*

Level 3. Elimination of duplicate semes: *syntactic dependency tree annotation =>*  *dependency tree annotation.*

Comparison of 2- and multicomponent NPs within a text gives evidence of 2-component NPs being the source for longer constructions. When two 2-component NPs result in a multicomponent structure within the boundaries of one text or texts of the same subject area, one can establish a number of patterns for the resulting combination:

* a 4-component NP as a result of merging two NPs of A + N type, embedding NP1 as an attribute before the head of NP2:

*seismic analysis+ indirect method ⇒ indirect seismic analysis method*

*second language + adult learner ⇒ adult second language learner*

1. a 3-component NP, when two initial NPs have a common component:

*mental processing + processing operation ⇒*

*mental processing operation*

1. a 3-component NP, when one component in NP1 is semantically supported by a component in NP2:

*communicative method + language learning ⇒ communicative language learning*

1. a 3-component NP, when the semantics of the resulting NP is determined by the domain extralinguistic knowledge, for example:

*seismic stability + direct analysis ⇒ seismic stability direct analysis*,

which in the text may be convoluted up to a 3-component NP *seismic direct analysis.*

1. **Translating multicomponent NPs as a part of lexicographic analysis**

NP standard transformations described in section 2 do not show all possible variants of NP development in a text. However, they might be helpful when translating an NP with a high degree of structure compression. As the research shows, it is exactly 2-component NPs that present particular difficulties in their analysis and translation.

To overcome the difficulties, we find only two approaches which can be used both in MT and human translation.

The first approach includes modelling the knowledge base of the domain in question (within the framework of the MT system) or appealing to the translator’s factual knowledge. In the case of MT this approach is based on extensive research into possible relationships between the basic concepts of the domain and the items of the linguistic database. Creating such a thesaurus or a semantic net is not only extremely laborious, but also space-consuming. And the most serious disadvantage is that sometimes it is impossible to achieve an unambiguous solution to the problem. For example, a semantic network for *constant amplitude deformation cycle* would show relations between the nodes *constant* and *amplitude*, *constant* and *deformation*, *constant* and *cycle,* and this information doesn’t help to establish the dependencies structure of the NP both in MT and human translation.

The second approach is more formal: we can use the information obtained from the analysis of the entire text. This approach seems more appropriate, since it is based on formal indicators of the author's intentions, which are reflected both in the text structure and in the composition of different NPs with the same components.

NP contextual analysis within the text space, provided by concordancing in scientific text corpora, leads to establishing procedures of coining novel NPs from those featuring in the text and to recognizing the compressed sentence structure in a concise form of an NP.

To establish the procedures, we suggest to use MT results for the source part of a parallel corpus as a reference base. Thus, comparing the translations of the English part (source) and the Russian part in a parallel corpus, we can find exact matches of NPs, as well as partly coincidence for term components and their presence in full and compressed terms.

For instance, a 3-component NP *design equipment models* in the source English part can be variably translated as *модели расчетного оборудования* or *расчетные модели оборудования.* The English part also has a 2-component NP *design models*, its MT is *расчетные модели,* which finds an exact match in the Russian part: *расчетная модель.* But there is no variant of *design equipment* withan expected MT *расчетное оборудование.* Nothing similar is found in the Russian part, either (see Table 3). The comparison suggests that *design models/расчетные модели* demonstrates stronger dependences between *design* and *models* in the texts of this subject area*,* than between *design* and *equipment.* So, the right candidate for a dictionary entry is *расчетные модели оборудования.*

*Table 3.*  **Comparing NPs in a three-part corpus (Subject Area** «Seismic Protection»)

|  |  |  |
| --- | --- | --- |
| **English part** | **MT stored results** | **Russian part** |
| *design models*  *design equipment models* | *расчетные модели*  *\*модели расчетного оборудования / расчетные модели оборудования* | *расчетная модель*  *конструктивная модель*  *проектная модель* |

Analysis of texts across different subject areas has shown, that if an NP of more than two components appears in the text, it is generally followed by a 2-component NP in the nearest context within the limits of 2-3 sentences, or it can be found in the title, keyword list or abstract. Hence, in human translation this fact can be a clue for NP structure diagnostics. Searching parallel corpora, we may fail to fix such relations, but referring to MT results as a storage base, we can optimize term identification and translation.

1. **Conclusion**

We have demonstrated that a scientific text abandons in multicomponent terminological NPs, most frequently 2-component combinations, with ambiguous dependency relations. This ambiguity is caused by their syntactic compression, since an NP is normally the result of a sentence or of another NP convolution.

To establish dependency relations of a multicomponent terminological NP it is useful to seek for its variants and modifications that can be found within the same text or texts of the subject area. These modifications are results of a number of standard transformations described in the paper.

Our study of various subject area texts has shown, that if an NP of more than two components appears in the text, it is generally followed by a 2-component NP in the nearest context, or it can be found in the title, keyword list or abstract.

While context analysis and comparison of NP modifications within a text may serve a reliable clue in human translation and manual/traditional lexicographic work, searching parallel and comparable text corpora for terminological equivalents may find few exact matches and the NP modifications may show no kinship of the components. We suggest to optimize the use of full-text parallel corpora for lexicographic and terminographic aims by adding a third part ̶ an MT results corpus ̶ as a reference base to fix and store the history of NP transformations and modifications in the corpus.

**References**

1. Babych, B., Hartley, A. (2003). Im­proving machine translation quality with automatic named entity recognition. In Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT. Association for Computational Linguistics, pp. 1–8.
2. Baroni, M., Zamparelli, R. (2010). Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space. In Proceedings of the 2010 Conference on Empiri­cal Methods in Natural Language Processing, pp. 1183–1193.
3. Beliaeva L. (2014). Applied Lexicography and Scientific Text Corpora // Transcations on Business and Engineering Intelligent Applications. Galina Setlak, Kassimir Markov (ed.). Rzeshov, Poland: ITHEA, pp. 55–63
4. Belyaeva, L. (2009). Scientific Text Corpora as a Lexicographic Source // SLOVKO 2009. NLP, Corpus Linguistics, Corpus Based Grammar Research, Proc. from the Intern.Conference, November 25 – 27 2009, Smolenice, Slovakia, pp. 19–25
5. Bergsma, S., Wang, Q.I. (2007). Learning noun phrase query segmentation. In Proc. EMNLP- CoNLL, pp. 819–826.
6. David, V., Curran, J. (2007). Adding noun phrase structure to the penn treebank. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguis­tics.Association for Computational Linguistics, Prague, Czech Republic, pp. 240–247.
7. Delgado, M., Martin-Bautista, M.J., Sanchez, D., Vila, M.A. (2002). Mining Text Data: Special Features and Patterns // Lecture Notes In Computer Science, Vol. 2442, Springer-Verlag GmbH, pp. 140–151
8. Delpech E., Daille B. (2010). Dealing with lexicon acquired from comparable corpora: validation and exchange // Proceedings, 9th Conference on Terminology and Knowledge Engineering (TKE). − Fiontar, Dublin City University, pp. 229–223.
9. Feldman, R., Dagan, I. (1995). Knowledge discovery in textual databases (KDT) // Proceedings of the 1st Int. Conference on Knowledge Discovery and Data Mining (KDD-95), AAAI Press, pp. 112–117.
10. Heja, E. (2010). The Role of Parallel Corpora in Bilingual Lexicography. In: N. Calzolari et al. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10). Valetta: European Language Resources Association (ELRA), pp. 2798–2805.
11. Lavie, A., Parlikar, A., Ambati, V. (2008). Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora. In Proc. 2nd SSST, Associa­tion for Computational Linguistics, pp. 87**–**95.
12. Malakhovskaya, M., Beliaeva, L., Kamshilova, O. (2021). Teaching Noun-Phrase Composition in EAP/ESP Context: A Corpus-Assisted Approach to Overcome a Didactic Gap // Journal of Teaching English for Specific and Academic Purposes, Vol. 9, No 2, pp. 257–266. DOI Number https://doi.org/10.22190/JTESAP2102257M.
13. Reiter, N., Frank, A. (2010). Identifying Generic Noun Phrases. In Proceedings of the 48th Annual Meeting of the Association for Computa­tional Linguistics**,** Uppsala, Sweden, July. Association for Computational Linguistics, pp. 40–49.
14. Shen, L., Xu, J., Weischedel, R. (2008). A new string-to-dependency machine translation algorithm with a target de­pendency language model. In Proceedings of ACL-08: HLT, pp. 577–585.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Беляева Лариса Николаевна**

Российский государственный педагогический университет им. А.И. Герцена (Россия).

**Beliaeva Larisa**

Herzen State Pedagogical University of Russia (Russia).

***E-mail: lauranbel@gmail.com***

**Камшилова Ольга Николаевна**

Российский государственный педагогический университет им. А.И. Герцена, Санкт-Петербургский университет технологий управления и экономики (Россия).

**Kamshilova Olga**

Herzen State Pedagogical University of Russia, Saint-Petersburg University of Management Technologies and Economics (Russia).

***E-mail: onkamshilova@gmail.com***

1. http://www.ttc-project.eu/about-ttc/concept-and-objectives [↑](#footnote-ref-1)
2. The «Seismic Protection» corpus a 1-million-word partly parallel corpus, the size of English and Russian parts is 500 000 tokens each (19145 English and 25872 Russian wordforms respectively) [↑](#footnote-ref-2)
3. The «Web and Linguistic Technologies» corpus is a comparable corpus, the English part including 372 English texts (3 468 000 tokens). [↑](#footnote-ref-3)