Nataliya Popovych, Andriy Lutskiv, Alla Tyshchuk
Uzhhorod – Ternopil', Ukraine

Corpus-Based Concept Translation

The paper makes an overview of the corpora and corpus tools which are classified according to their content, functionality, aim and adaptability. The adaptable corpus tool is developed to complete specific linguistic tasks and aimed at resolving the issues of concept translatability while doing comparative corpus-based translation studies research. Main tasks of translating concepts are discussed and some examples of adaptable corpus-based comparative concept analysis are shown. The linguistic background methodology of the concept translation study is based on conceptual analysis, componential analysis, semantic triangle theories, system of values theory.

Keywords: concept translation, religious concept, corpus-based translation studies, conceptual analysis, system of values

There are a great number of corpus tools and corpora applied in interpreting and translating. Such corpus-based approach has much to do with the context-based approach in language translating. Using corpus as a research tool a researcher has every possibility to study words, collocations, grammatical forms etc. directly from the context of the natural language he or she studies, to assess language unit usage, its frequency and stylistic peculiarities. Corpus-based translating is not only text- and context-oriented. Nowadays it tends to be a corpus-based and concept-oriented approach.

The methodological apparatus of this paper is based on such approaches to concept study and concept translation as conceptual analysis, componential analysis, religious concept study, semantic triangle theories, system of values theory. Corpus-based conceptual analysis of the concepts was realized by means of an adaptable corpus tool developed for completing specific linguistic tasks [23].

Corpora Classification: Overview. We divide corpora and corpus tools into groups: content-based, functionality-based, aim- or purpose-based and generation-based corpora and corpus tools. A great number of different linguistic tools, i.e., so-called corpus software tools are directed toward the accomplishment of one task, either linguistic or statistic in its nature. Among them are offline and web-based concordancers like AntConc (v.3.5.8, February 18, 2019) [7], WordSmith Tools (v. 7, 2019) [29], #LancsBox (v 4.0, 2018) [9], JConcorder (ver. 1.beta.13, 2011) [27], text coding, (manual) annotation programs, text-analysis tools & search engines like DART (ver. 3.0, 2019) [34], Dexter [14] and tools & resources for transcribing, annotating or analyzing texts (inc. speech or audio-visual) like CLaRK, ELAN (EUDICO Linguistic Annotator), GATE (General Architecture for Text Engineering), stats tools like Log-likelihood and effect size calculator, taggers like CLAWS, Stanford POS tagger and others [23, p. 217-218].

Content-based corpora and corpus tools can be subdivided into 1.1. national, 1.2. professional, 1.3. parallel, 1.4. comparable, 1.5. specialized and 1.6. task-based (adaptable or mixed). The Brown Corpus of Standard American English or the Brown Corpus by W.N. Francis and H. Kucera, the British National Corpus managed by the BNC Consortium and the Corpus of Contemporary American English are the most vivid examples of national corpora of both British and American English.

Ukrainian National Corpora are represented by several projects which have been realized till nowadays. Corpus of the Ukrainian Language (N. Dartchuk, O. Siruk, M. Langenbach, Ya. Khodakivska, V. Sorokin at the Institute of Philology of TKU of Kyiv), Laboratory of Ukrainian (Ukrainian) [22] and General Regionally Annotated Corpus of Ukrainian (GRAC) (Ukrainian) [13], are the most developed Ukrainian language corpora and corpus tools. English corpora and corpus tools have larger range of choice and are of different content, purpose and functional capacity.

Professional corpora and corpus tools are more content-oriented and focused on specialist language and vocabulary, like Air Traffic Control (ATC) Corpus [16] or Carnegie Mellon Communicator Corpus [8].

OPUS is one of the best examples of parallel multilingual corpora which contains converted and aligned free online data and added linguistic annotation. OPUS project team provides the community with a publicly available parallel corpus. It is based on open source products and the corpus is also delivered as an open content package [23].

Another type of multilingual corpus is comparable corpus which consists of original texts rather than translations where all texts are similar in content, but they differ in languages or language varieties in the sense that the texts of the same domain are aligned [18]. These types of corpora are aimed at comparing the languages or varieties presented in similar circumstances of communication, without the distortions which appear in translated texts of parallel corpora [18].

To specialized corpora belong BASE (British Academic Spoken English) compiled by Hilary Nesi and Paul Thompson, BAWE (British Academic Written English), LANCAWE (Lancaster Corpus of Academic Written English), to name just a few.

All above named corpus groups can be also classified according to their functional annotation set into linguistic on the word level, syntactic, semantic and discourse. According to their aim or purpose corpora or corpus tools are divided into corpora for linguistic research and statistical data extraction.

McEnery and Hardie and then L. Anthony outline four generations of corpora available today. The first generation appeared in the 1960s and 1970s and ran on mainframe computers, were able to process the ASCII character set and were limited to processing only English corpora. The advantage of the second generation of corpus tools is that they could run on the early personal computers, allowing researchers to carry out small-scale studies and allowed teachers to introduce corpora analyses into the language learning classroom, i.e. Data-Driven Learning (DDL) approach. The third generation corpus tools struggle to handle very large corpora of over 100 million words. Automatically compiled by scraping data from Internet sites, these corpora can be several billion words in length, and the architecture of third generation tools is not appropriate to process them. Hence, they cannot be used for analysis with corpus tools on a personal computer. The fourth generation tools, such as corpus.byu.edu (Davies 2013), CQPweb (Hardie 2013), SketchEngine (Kilgariff 2013), and Wmatrix (Rayson 2013) are tools which offer better scalability by storing the corpus in a Web server database and pre-indexing the data. Contemporary corpus tools have functions to analyze KWIC concordances, distribution plots, clusters and N-grams, collocates, word frequencies, and keywords. Most of the tools are still English-centric in that they only allow access to English corpora, which is a great disadvantage for Ukrainian corpus users [6].

Methodological Apparatus for Completing Concept Translation. M. Kosterec, R.Jackendoff, J.Horvath, I. Dahlberg, A. Nuopponen and many other linguists and philosophers focused on conceptual analysis application in its different form of use and domains. Ch. Stead, R.E. Witt, G.Dörrie, V.H. Drecoll, G.-L.Prestige, M. Simonetti, D. Spada, O. Biletskyi, A. Biletskyi, S. Averyntsev, N. Saharda, V. Bolotov, J.N.D. Kelly, G. Reale, Ch. Yannaras focused on religious concepts, especially on those used in the texts of the Golden Age of Patristics and studied the connections between them on lexico-semantic level.

Componential analysis is a well-known linguistic approach to semantic meaning study which originated in the works of F.G. Lounsbury and W.H. Goodenough on kinship terms and was further developed by O.K. Seliverstova, J.N. Karaulov, E. Nida, D. Bolinger and other linguists.

The origin of the semantic triangle theories can be traced back to the 4th century BC in Aristotle's *Peri Hermeneias* in its Latin translation *De Interpretatione*, i.e., the second book of his *Organon*.

I.V. Arnold says that originally this triangular scheme was suggested by the German mathematician and philosopher Gottlieb Frege (1848-1925). It found its future applicability in the work of the English scholars C.K. Ogden and I.A. Richards in the form of triangle of reference and was transformed into the theory of semantic triangle by other linguists like F. de Saussure and others.

System of values theory by R. Jackendoff deals with the conceptualization of values and how humans conceptualize them in different religious traditions, cultures, social groups etc.

The notion of "conceptual analysis" was used by many researchers and applied in different linguistic domains. It is also regarded to be an ambiguous term due to the fact that there is no exact definition and unique application of conceptual analysis in linguistics.

In translating concepts there is a set of main issues a translator or interpreter has to deal with: (1) so-called transcoding of the source language concept, (2) corpus-based conceptual analysis, (3) achieving high level of translation quality, (4) reaching conceptual equivalence in translation. Apart of using CAT tool of a certain kind to facilitate the process of translation, especially for larger texts, a translator will feel more comfortable while using also some corpus tools to check the concept under analysis on the level of national language text collections. An adaptable corpus tool has been developed to meet the needs of translators depending on the translating tasks especially when dealing with religious or historical concepts.

Transcoding of the Source Language Concept or Semantic Triangle Theories. The representation of the concept understanding in the form of triangle has its long history. To introduce a new element into this theory, it is worth mentioning three triangle theories which are of the utmost importance for this research. Semantic triangle by Ferdinand de Saussure, C.K. Ogden and I.A. Richards' Triangle of reference, Arnold's triangle of meaning are the vivid representations of triangle relationship that forms the basis for the understanding of the concept. Triangle of value or conceptual triangle results from the system of values theory because of its crucial importance in the understanding of concept development process in the history of language. The system of values theory was developed by Ray Jackendoff, the founder of conceptual semantics. The understanding of concept translating directly depends on "how humans conceptualize systems of value. Value can be thought of as an abstract property attributed to objects, persons, and actions. There are several distinct types of value, i.e., affective value, utility, normative value, personal normative value, and esteem. Values also can be differentiated as subjective verses objective. Several important inferences drive the interaction of multiple values in determining one's course of action and one's expectations of others' actions. These are reflected in our understanding of such notions as fairness, reciprocity, restitution, honoring, shaming, and apology" [17].

According to R. Jackendoff, value is a conceptualized abstract property attributed to conceptualized objects, persons and actions where value can be equal to word (symbol) and inference to referent in the semantic triangle theories (Fig.1). Referent is the same as concrete lexical meaning represented by a concrete object or abstract notions. "The values of an entity play the role in rules of inference that affect the ways one reasons about the entity"[16].

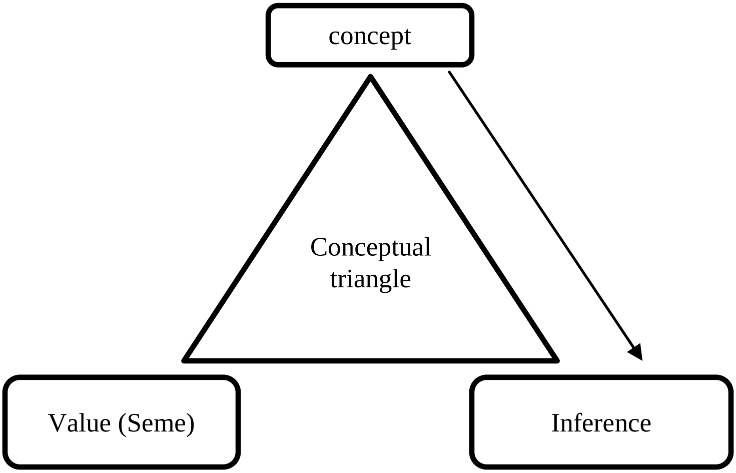

Fig.1. Conceptual Triangle

Corpus-Based Conceptual Analysis. The question of the equivalent / adequate concept translation is quite a challenging one. It requires the understanding of what other subfields of linguistics tell us about concept and which of their theories and approaches are to be taken into account. The project of adaptable corpus development is on its initial stage and as an interdisciplinary approach to concept analysis it applies linguistic, mathematical and IT methods. It is aimed at examining concepts and their connections and interconnections in the source language text and its translations into several languages.

Corpus-based translation studies analysis of the concept is based on interdisciplinary methodology. From the one side such notions as concept relations, concept systems and the place of conceptual seme in the system of concepts are taken into account. From the other side, it is possible only by means of some additional corpus-based tool named adaptable text corpus tool able to analyze big amounts of text data [23].

Corpus-based concept analysis is realizable thanks to the employment of LSA and SVD approaches. "Latent semantic analysis (LSA) is a technique in natural language processing and information retrieval that seeks to better understand a corpus of documents and the relationships between the words in those documents. It attempts to distill the corpus into a set of relevant concepts. Each concept captures a thread of variation in the data and often corresponds to a topic that the corpus discusses. Without yet delving into the mathematics, each concept consists of three attributes: a level of affinity for each document in the corpus, a level of affinity for each term in the corpus, and an importance score reflecting how useful the concept is in describing variance in the data set" [28, p. 115].

LSA as a technique in natural language processing, which is based on distributional semantics in particular, analyzes "relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text (the distributional hypothesis)" [31]. It comprises four main steps: Term-Document Matrix, transformed Term-Document Matrix, Dimension Reduction and Retrieval in Reduced Space [31, p.194].

Singular value decomposition (SVD) is used at step 3 to reduce the number of rows while preserving the similarity structure among columns.

SVD "starts with a document-term matrix generated through counting word frequencies for each document. In this matrix, each document corresponds to a column, each term corresponds to a row, and each element represents the importance of a word to a document. SVD then factorizes this matrix into three matrices, one of which expresses concepts in regard to documents, one of which expresses concepts in regard to terms, and one of which contains the importance for each concept" [28, p. 115].

Every concept as such belongs to the system of concepts and has its relations within the system of concepts. By means of LSA and SVD approaches it is possible to analyze connections between concepts on the example of huge amount of text data, which is very useful for identifying the meaning of the concepts, semantic similarities etc. in comparative translation studies analysis. A minimal conceptual seme has its important functional role in that big system of concepts and their relations.

As A. Nuopponen admits, "Concept relations and concept systems are inseparable, since without relationships there would be no system, and since relationships depend on the systemic context. Concept relations may be strictly logical connections or freer associations between one concept and another. They are mental entities which link concepts to one another. Concept relations are thus one type of concept, concepts of relationship, and, like other concepts, they are the result of abstraction. Their referents are the relations between individual entities, whether it is a question of similarity or other relations" [22]. Table 1 shows results of finding top 10 related terms to concept "light", «світло», «свет» on the example of 1189 texts of the Bible and reveals the possible 10 conceptual semes of "life", «життя», «жизнь».

Text	**Num.**	**Results of finding top 10 related terms to term "life", «життя», «жизнь»**
[10]	1189	[life 0.9999999999999999], [way 0.7831508394861265], [slander 0.7762407019125012], [love 0.7756090309091765], [keep 0.7609397876061197], [deceit 0.7271818031322735], [always 0.726626405138741], [eye 0.7203850472552942], [teach 0.711317467929176], [learn 0.7102973298476237]
[18]	1189	[life 0.9999999999999998], [slander 0.7906746918634807], [way 0.7644799466537877], [eternal 0.752632237317999], [gain 0.7493386275750088], [good 0.7473726160229541], [keep 0.7460081819756795], [test 0.731392381840476], [consider 0.7304389665215004], [love 0.7291446846225138]
[17]	1336	[life 1.0], [good 0.7896825311885981], [beginning 0.7604608631773508], [man 0.7566270752300552], [know 0.740360584555423], [knoweth 0.7368168194109597], [doth 0.7330186673322208], [loveth 0.7256017136728831], [nothing 0.7254483035496164], [thing 0.7252873055182654]
[4]	1189	[життя 0.9999999999999994], [людина 0.7782014335272919], [мати 0.7620087738950805], [слово 0.7573906349775805], [серце 0.7247618568385172], [бачити 0.7065606382250862], [добрий 0.7063032316840142], [смерть 0.7035770056981053], [бог 0.7029945213472155], [знати 0.7010595042140018]
[2]	1189	[життя 0.9999999999999998], [мати 0.7752990316095201], [людина 0.7544012004308451], [слово 0.7267753377907553], [знати 0.7263722705993478], [смерть 0.724361094302647], [серце 0.7116854979756364], [бачити 0.6977814747163861], [бог 0.6883015389649128], [добрий 0.6811600046127868]
[1]	1189	[жизнь 1.0], [человек 0.820781169627828], [сердце 0.7951173382032621], [бог 0.7860838875388204], [длить 0.7716773924299442], [слово 0.7651749438040882], [мир 0.7634569020514778], [мень 0.7620893832639917], [миро 0.7593554273284232], [тома 0.7504558415796446]

References

1. Библия. Новый русский современный перевод «Слово Жизни». Международное Библейское Общество, изд. "Biblica". 2014. 998 с.
2. Біблія або Книги Святого Письма Старого й Нового Заповіту. Із мови давньоєврейської й грецької на українську дослівно наново перекладена. [переклад професора І. Огієнко]. К.: УБТ, 1962. 1529 с.
3. Біблія (четвертий повний переклад з давньогрецької мови), переклад ієромонаха о. Рафаїла (Романа Турконяка). Львів: Українське Біблійне Товариство, 2011. [Електронний ресурс] ukrbible.at.ua/load/zavantazhiti_ukrajinsku_bibliju/skachaty.../7-1-0-165.
4. Новітній переклад Біблії Олександра Гижі. К.: Друк КТ Забеліна-Фільковська, 2013. 1210 с.

5. Святе Письмо Старого і Нового Завіту. Мовою русько-українською: переклад П.О. Куліша, І.С. Нечуя-Левицького, І. Пулюя. К.: Простір, 2010. 852+249 с. **6**. Anthony L. A critical look at software tools in corpus linguistics, (2013) Linguistic Research, 30 (2), pp. 141-161. https://doi.org/ 10.17250/khisli.30.2.201308.001. **7**. Anthony L. AntConc (Windows, Macintosh OS X, and Linux) [Online]. Available: http://www. laurenceanthony.net/ software/antconc/releases/AntConc358/help.pdf. **8**. Bennett Ch., Rudnicky A.I. The Carnegie Mellon Communicator Corpus, in Proc. of the Internat. Conf. of Spoken Language Processing, Denver, Colorado, 2002, P.341-344. **9**. Brezina V., Timperley M.& McEnery T. (2018). #LancsBox v. 4.x [software]. [Online]. Available: http://corpora.lancs.ac.uk/lancsbox. **10**. Christian Standard Bible. Holman Bible Publishers, 2018. 1920 p. **11**. Comparable Corpora [Online]. Available: https://www1.essex.ac.uk/linguistics/external/clmt/ w3c/corpus_ling/content/corpora/types/comparable.html/. **12**. Corpus Finder. [Online]. Available: http://www.helsinki.fi/varieng/CoRD/corpora/ corpusfinder/. **13**. Corpus of the Ukrainian Language (Ukrainian) [Online]. Available: http://www.mova.info/corpus.aspx?l1=209. **14**. Garretson G., Dexter. Tool for analyzing language data [Online]. Available: http://www.dextercoder.org/ index.html. **15**. General Regionally Annotated Corpus of Ukrainian (GRAC) (Ukrainian) [Online]. Available: http://uacorpus.org/, http://www. parasolcorpus.org/bonito/run.cgi/first_form. **16**. Godfrey J.J. Air Traffic Control Complete [Online]. Available: https://catalog.ldc.upenn.edu/ LDC94S14A. **17**. Holy Bible: King James Version, Study Edition, Containing The Old Testament, Apocrypha, and New Testament / American Bible Society//2011. 1462 p. **18**. Holy Bible. New International Version. Zondervan Publishing House (Grand Rapids, Mich.) 2011. 1140 p. **19**. Jackendoff R. On conceptual semantics. *Intercultural Pragmatics*. 2006. 01 Vol. 3; Iss. 3. Walter de Gruyter GmbH & Co. KG. P.353-358. **20**. Jackendoff R. The Peculiar Logic of Human Values (Lecture). Santa Fe Institute (26 April 2012). [Online]. Available: https://www.youtube.com/ watch?v=vRc3MiP6Tok&t=1187s . **21**. Kübler S., Zinsmeister H. Corpus Linguistics and Linguistically Annotated Corpora (English). New York; London: Bloomsbury Academic, 2015. P. 21-156. http://dx.doi.org/10.5040/ 9781472593573. **22**. Laboratory of Ukrainian (Ukrainian) [Online]. Available: https://mova.institute/. **23**. Lutskiv A., Popovych N. Adaptable Text Corpus Development for Specific Linguistic Research. *Proc. of 2019 IEEE International Scientific and Practical Conference "Problems of Infocommunications. Science and Technology"*. 08-11 October 2019, Kyiv, Ukraine, IEEE Catalog Number: CFP19PIA-USB; ISBN: 978-1-7281-4183-1. P. 217-223. **24**. Nida E.A. Towards a science of translating. With special reference to principles and procedures involved in Bible translating. Leiden : Brill, 1964. 331 p. **25**. Nuopponen A. Begreppssystem för terminologisk analysis (Concept systems for terminological analysis). *Acta Wasaensia*. No 38. 1994. 266 p. **26**. Opus, the open parallel corpus [Online]. Available: http://opus.nlpl.eu/. **27**. Rand D. JConcorder. Java-based Concordance software [software]. [Online]. Available: http://www. concorder.ca/index_en.html. **28**. Ryza S., Laserson U., Owen S., and Wills J. Advanced Analytics with Spark. Patterns for Learning from Data at Scale.

2nd ed. Sebastopol: O'Reilly Media, Inc., 2017. P. 115-136. **29**. Scott M. WordSmith Tools Manual [Online]. Available: https://lexically.net/downloads/version7/HTML/index.html. **30**. Standard American English or the Brown Corpus [Online]. Available: http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM#t1. **31**. Susan T. Dumais. Latent Semantic Analysis. *Annual Review of Information Science and Technology*. 38: 2005. P. 188-230. doi:10.1002/aris.1440380105. **32**. The UCREL semantic analysis system https://www.researchgate.net/publication/228881331_The_UCREL_semantic_analysis_system. **33**. Rajaraman A, Leskovec J., Ullman J.D. Mining of Massive Datasets. [Online]. Available: http://infolab.stanford.edu/~ullman/mmds/book.pdf. **34**. Weisser M. Manual for the Dialogue Annotation & Research Tool (DART), Version 3.0 [Online]. Available: http://martinweisser.org/publications/DART_manual_v3.0.pdf. **35**. Weisser M. Corpus-based Linguistics – Introduction, 2016; last edited: Wed Apr 03 14:08:29 2019, [Online]. Available: http://martinweisser.org/corpora_site/CBLLinks.html.

Barbora Tomečková
Brno, Czech Republic

The Effects of Defects

Mechanical translators are in the 21st century an inherent part of translatology. From using the old-school camera and paper cards the technology has shifted to the use of the neural networks, which means the need of technology to use statistical machine translation or neural networks to deep learning. The main translators which started using neural networks are Google Translator, Bing Microsoft Translator and Facebook´s automatic translation service [1].The technology is using its own brain to learn similarity between languages and evaluate the probability of translations. In contrary, the neural machine translation is not perfect. The proof are funny pictures on the Internet, but not every mistake translators made is that banal.

Keywords: translation, translate, translator, issues, defects, neural networks, mechanical translation

History of translators

The beginning

The history began in 1933. At the time the Soviet scientist Peter Troyanskii created the machine for the selection and printing of words when translating from one language to another. He only used cards in four different languages, a typewriter and an old-school film camera [2]. He selected a word and printed it when he was translating from one language to another one [3]. The person who was using it typed a word, took a photo and wrote some information about the word. The film made frames with