Information Structure of Contemporary Popular Scientific and Technical Text

Larisa Ilinska, Marina Platonova, Tatjana Smirnova

Faculty of E-Learning Technologies and Humanities, Riga Technical University

Structure

- Diverse nature of Information;
- Information structure and dichotomies;
- Foregrounding;
- Information processing;
- Formalization degree;
- Information extraction;
- CAT tools;
- Conclusion.

Information

knowledge

meaning

comprehension

constraint

perception

representation

communication

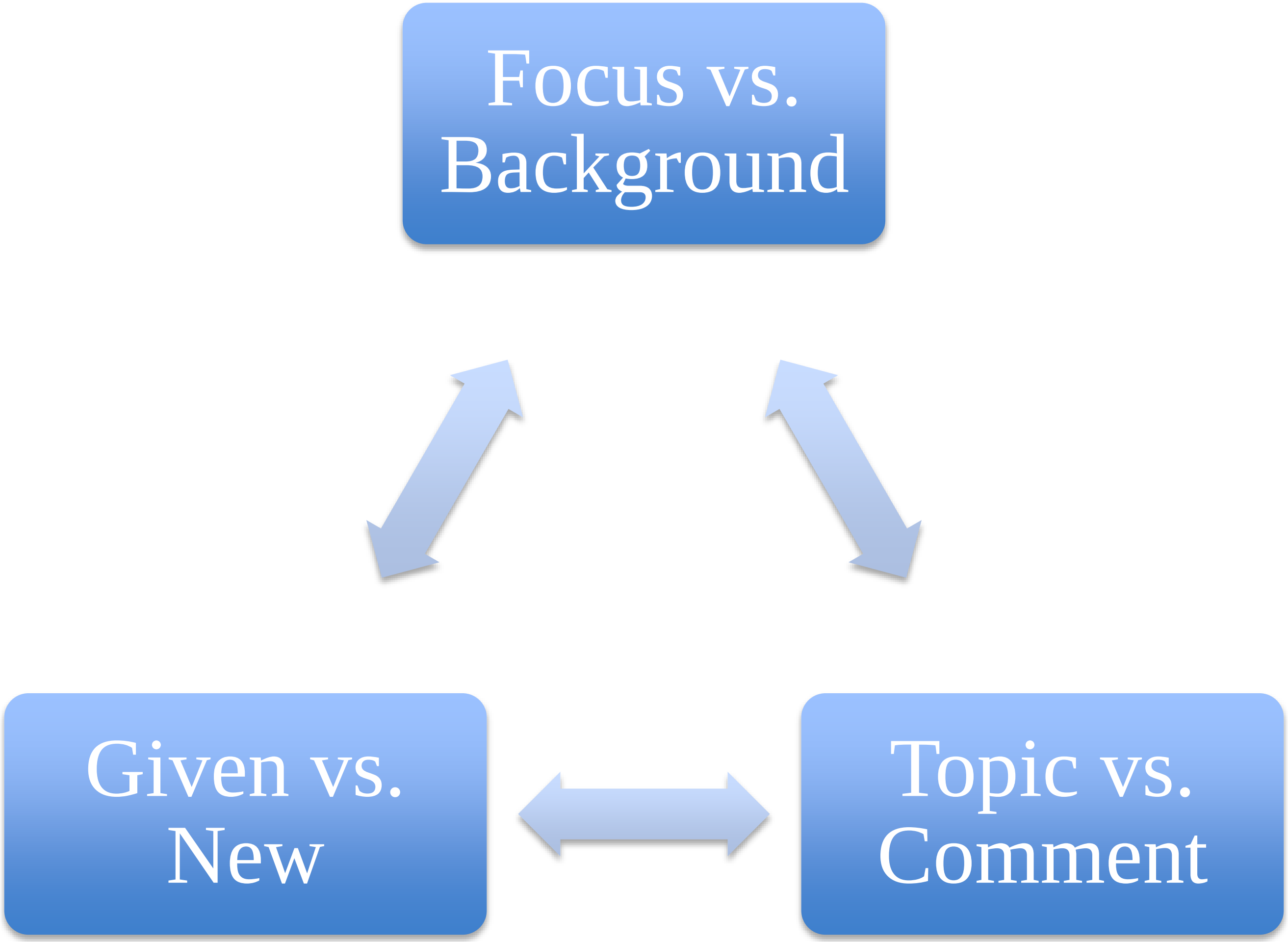
Information Structure

Investigations on the information structure of scientific and technical texts have become particularly topical with the introduction of new methods of text analysis using corpora and text processing software.

Information structure

Scientist	Definition
Lambrecht	the formal expression of the pragmatic structuring of a proposition in discourse
Schwabe and Winkler	the term Information Structure refers to the linguistic encoding of notions such as focus versus background and topic versus comment, which are used to describe the information flow with respect to discourse-givenness and states of activation

Information Structure Dichotomies



Focus vs. Background

Given vs. New

Topic vs. Comment

Foregrounding

New information is often brought into focus using various foregrounding techniques such as application of metaphoric terms, allusions, proverbs, idioms, and terms belonging to different fields of knowledge.

Foregrounding

The application of stylistically marked vocabulary within the scientific and technical text allows focusing the attention of the readers on a particular information cluster, ensuring that the new information is not disregarded or missed.

Information Processing

The challenges associated with processing of information and its extraction from the text are rooted in the fact that even the most advanced computer-aided text processing methods are incapable of performing many tasks unless they are combined with the methods of cognitive analysis

Degree of Formalization of Language

Positive tendency	Drawback
It facilitates the process of data mining and information extraction, because of the clear information architecture (certain order of data representation) and pre-defined set of representative features (stated tokens)	It becomes more difficult to adjust the text processing software to any changes in the order of the given information, which does not make the data mining and information extraction systems as flexible as possible to new situations.

Information Extraction: Interlingual Setting

- Disability to compensate the loss of stylistic coloring when aligning a metaphoric term into a more formal language,
- Disability to generate a metaphoric term based on the associative mining function to be used in the less formal language
- Disability to assign meaning to linguistic expression taking into account the existing information, i.e. micro- and macro-context

CAT Tools

Modern text processing tools should be able to perform multiple tasks,

- classifying texts according to genres and functions,
- distinguishing intra-disciplinary and cross-disciplinary polysemic terms/words,
- decoding different models of meaning extension, and culture-specific items.

CAT Tools: Potential Challenges

Linguistic phenomenon	IT challenge
Tendency to present information implicitly	Challenges in decoding and translating sender’s implicatures and presuppositions
Tendency for uncontrolled metaphoric meaning extension of the existing lexical items	Challenges in identifying, tracing and extracting metaphoric (hidden, covert, connotative) meaning of the ad hoc created unlexicalised metaphoric lexical items
Appearance of polysemic terms	Challenges in differentiating and employing terms regarding its status and the context of application
Appearance of occasionalisms and elements of professional jargon	Such items of professional vocabulary are not fully lexicalized and, as a result, are not always recognized by CAT tools

CAT Tools

The term “body” as used in the field of chemistry may be presented as follows:

- consistency,
- saturation,
- coverage capacity,
- strength,
- proof,
- viscosity,
- density,
- thickness,
- extractivity,
- intensity,
- glutinosity

Conclusion

The information structure of the contemporary popular scientific and technical text is characterized by the distinct hierarchical organization, growing information density, and the increased degree of intertextuality, i.e. interaction between the given and new information.

Conclusion

Natural language is characterized by uncontrolled creative use of language resources resulting in the infinite number of meaning combinations.

Conclusion

IE is complicated due to the presence of terms based on metaphoric meaning extension, proper names based on metonymy, intra-disciplinary and cross-disciplinary polysemy, and culture-specific items.

Conclusion

The challenges associated with decoding of meaning of foregrounded elements are most apparent when these elements should be communicated across the languages and recorded in multilingual databases.

References

- Govindarajulu N. S., Bringsjord, S., Licato, J.: On Deep Computational Formalization of Natural Language. Presented at Formalizing Mechanisms for Artificial General Intelligence and Cognition, FORMAL MAGIC 2013. Beijing, China.
- Hobbs, J.: The Generic Information Extraction System. In: Proceedings of the 5th Message Understanding Conference (MUC-5) (1993).
- Lambrecht, K.: Information Structure and Sentence Form: Topic, Focus, and the Mental Representation of Discourse Referents. CUP, UK (1994), cited p.5.
- Piskorski, J., Jangarber R.: Information Extraction: Past, Present and Future. In: Poibeau, T. Saggion, H., Piskorsi, J., Jangarber R. (eds.) Multi-source, Multilingual Information Extraction and Summarization, pp. 23-49. Springer-Verlag, Berlin (2013).
- Schwabe, K., Winkler, S.: On Information Structure, Meaning and Form: Generalization Across Languages. John Benjamins Publishing, Amsterdam (2007), p.1.
- Shannon, C. E., Weaver, W.: The Mathematical Theory of Communication. Foreword by Richard E. Blahut and Bruce Hajek. University of Illinois Press, Urbana (reprinted in 1998).
- Steube, A.: Information Structure: Theoretical and Empirical Aspects. Walter de Gruyter, Berlin (2004), cited pp.15-16.
- Turmo, J., Ageno, A., Catala, N.: Adaptive Information Extraction. ACM Computing Surveys, Vol. 38, 2, Article 4 (2006), cited pp.1, 2, 12.

Thank you for attention!