

Materials of Conferences

**THE ROLE OF COREFERENCE
IN NOMINATIVE PHRASES
IN OVERCOMING COMMUNICATION
FAILURES IN CHEMISTRY-ACQUISITION
IN BILINGUAL CONTEXTS**

Davydenko L.G., Butenko L.I.

*Pyatigorsk State Linguistic University, Pyatigorsk;
Pyatigorsk State Pharmaceutical Academy, Pyatigorsk,
e-mail: linguist_07@mail.ru*

The goal of our work is studying coreferent relations in nominative phrases (NPs) in order to overcome communication failures in chemistry-acquisition in bilingual contexts

Transnationalization of education aims to increase the percentage of overseas students in Russian higher education institutions, and this requirement is caused by the demand to improve their competitive potential.

The goal of our work is studying coreferent relations in nominative phrases (NPs) in order to overcome communication failures in chemistry-acquisition in bilingual contexts.

In Pyatigorsk Pharmaceutical Academy there are 85 international students from Asia, Africa and Commonwealth of Independent States. Sorry to say, in some of these countries the level of education does not quite meet Russian standards.

The profession of pharmacy is known to blend science, technical art and human relationships in a unique fashion. The subject of pharmacy is studying chemical substances. On the whole, basic to the science in pharmacy are contributions from biology and chemistry. As the teaching language is Russian, lecturers of Russian higher education face two principle problems: the language barrier and lack of specific knowledge in the subjects at the beginners' level.

It should be notified, that chemistry refers to a group of disciplines whose function is to adopt a system of student academic competence to explain the basic processes occurring in the real world. Besides, it introduces students to scientific knowledge. Making up a holistic chemical world picture in the minds of our students becomes possible only at a certain point of sufficient development of their cognitive interest, cognitive activity and successful communication. To optimize the process of successful communication in teaching the basics of chemistry to international students, it is necessary to correctly decode the key concepts embodied in conceptual systems of the chemical picture of the world.

A cognitive approach to studying the real world concepts means striving to understand this world, expanding information about the reality and in such a way finding expanded solutions to the problems a human being is facing.

In the process of teaching lecturers inevitably resort to the act of reference, and in this respect they should be very careful and precise working with international students.

We argue that overcoming communication failures in the process of teaching chemistry to international students in medical institutions is much more successful in case coreference resolution systems are thoroughly studied and the results are taken into consideration. We view reference resolution as a clustering task and hereby distinguish between hypernyms and the corresponding hyponyms.

At the first stage of teaching chemistry *coreferential communication* can be turned to only with the purpose of denoting anaphoric relation. According to M. Poesio, the term «coreference» is used to indicate both the annotation of (generalized) anaphoric information and of information about reference proper (Poesio, 2000). The term «anaphoric relation» is used to indicate the relation between two textual elements that denote the same object; the subsequent mention of an entity already introduced is often marked by means of a particular type of noun phrase (NP) called an anaphoric expression. It is used by a human speaker to avoid repetition referring to the same unity. Annotating corpora with information about such relations between elements of a text is useful both from a linguistic point of view and for applications such as information extraction.

Later on, it is possible to use *coreference descriptors* explaining their significance.

In our opinions, to establish coreference correlations, a lecturer should keep to two essential rules:

- distinct differentiation between hypernyms (linear primary coreference and the corresponding hyponyms (linear secondary co-reference);
- determination of the cluster radius, i.e. measuring the distance between two noun phrases.

Overcoming communication failures in the process of teaching the basics of chemistry to international students will be successful if we correctly make up coreference clusters to distinguish between hypernyms and hyponyms. Just as in *linguistics*, in *chemistry* we interpret a hypernym as a word (in our case – NP) with a general meaning that has basically the same meaning of a more specific word (NP) representing the relation of class to subclass. Hyponyms are NPs whose *semantic field* is included within that of another NP. In simpler terms, a hyponym shares a type-of relationship with its hypernym and is used as a descriptor reflecting any specific characteristics of the main concept.

Let us consider the hypernym «alkenes». It names the class of organic compounds. Its hyponym, «saturated hydrocarbons», is used to reveal the characteristic chemical properties of alkenes, emphasizing that in alkenes molecules all the ties are saturated «to the limit».

Another hyponym for alkenes, homologues of methane, is used to make an emphasis on its similarities to methane as far as its composition and characteristics are concerned.

These characteristics should be collected in conceptors i.e., synonymous multitudes, to explain to students that all of them are characteristics of the same concepts and they are only different in definitions. A combination of a cognitive approach of symbolic (sign) systems connected with their related concepts (conceptors) leads to the adequate understanding of the text.

Nowadays thesauri are worked out. Making up a thesaurus that can describe the lexicon of the language in its diversity and wholeness can expand the boundaries of knowledge. Knowledge classification can be, and often is, TAXONOMIC (sometimes called «entity classification») like the classification of chemical elements (which means that they are going to list one concept in one place only in the classification structure). The aspect of classification is of crucial importance for information retrieval as it helps to establish the context in which one concept or phenomenon might be studied within the document.

General (special as well) bibliographic classifications are existing systems with big vocabularies. These systems give provision to describe not only subject, but also the form in which it is presented, the time and place that subject is connected with, the language it is presented in the document, the physical quality of the carrier etc.

Some of these classification schemes have a hierarchical structure, some list both single and composed concepts and are basically enumerating all possible subjects predicted to be studied in the documents. Some, however, tend to have faceted feature, i.e. to be synthetic, enabling expression of an infinite number of subject combinations in the documents. These classification systems are widely accepted and used in hundreds of countries and translated into many languages.

It is important that new concepts are constantly being added to follow the growth of knowledge. The above mentioned schemes are available in the electronic form as well.

As they use symbols rather than words they are especially suitable for the multilingual environment of the Internet. They can be used as the basis for developing thesauri or for building and tailoring a list of indexing terms for specific purposes. They can be used to describe any object not just textual. Classification can be designed and suitable for information retrieval and in most European countries have rich traditions in using classification as a language independent indexing tool.

In Russia there are a few best-known thesauri widely used by chemists. The Thesaurus of Descriptors of Chemistry and Chemical Industries (ТЭХИМ) represents a set of descriptors indicating semantic relationships between them, covering

specific areas of chemistry, e.g. the foundations of organic chemistry. There are 5.373 keywords and 10.133 descriptors in it. In the focus of another well-known Thesaurus there are words of chemical terminology in 19 languages [ГОСТ 7.24-2007].

It is commonly observed that people avoid repetition by using a variety of noun phrases to refer to the same entity, and some human audiences (e.g. international students) can have difficulty in this respect. Students should be taught to view the problem as one of partitioning, or clustering, the noun phrases, and define each group of coreferent noun phrases as an equivalence class.

In the process of chemistry-acquisition communication failures occur when the clustering algorithm gets broken down as a learning problem, i.e. while extracting wrong descriptors from thesauri students make up incompatible NPs.

The lecturer's task is to explain the students the algorithm of extracting descriptors and to revise the meaning of the concepts mapping a collection of noun phrases onto the same entity in the Thesaurus.

According to Cardie and Wagstaff, the clustering approach has a number of important advantages over existing learning and non-learning methods for coreference resolution. The most important for our paper is the following: the clustering approach provides a flexible mechanism for coordinating context-independent and context-dependent coreference constraints and preferences for partitioning noun phrases into coreference equivalence classes (Cardie and Wagstaff, 1999, 82).

In machine word processing, clustering requires additional filters, which determine the threshold of the clustering radius. It is very important because all of the NPs used to describe a specific concept will be «near» or related in some way, i.e. their «conceptual distance» will be small. A description of each NP and a method for «measuring» the distance between two noun phrases, a clustering algorithm, can then group NPs together: NPs with distance greater than a clustering radius are not placed into the same partition and so are not considered coreferent.

Lecturers, on the contrary, should determine the clustering radius and hereby coreference links intuitively via a set of hand-crafted heuristics and filters.

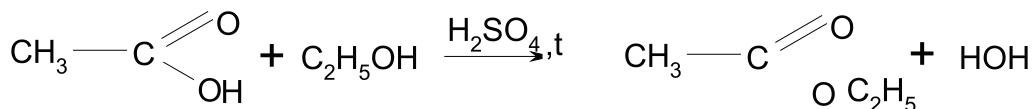
The principle characteristic features of coreferential links are as follows:

- Lexical features, i.e. the use of proper names and pronouns. E.g., in the class of alkenes in chemistry the names of separate members (homologues) act as proper names (Methane, Ethane, etc.). In more complicated cases, substances are given names according to certain rules, nomenclature – UPAC. Students should be explained these rules, which themselves are coreferentially independent.

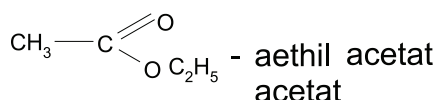
- Grammatical features, i.e. the features testing the grammatical properties of one or both of NPs (Ng, 2007, 1692).

• Semantic features. There are two semantic features, both of which are employed by Soon, Ng, and Lim's coreference system (Soon, 2001).

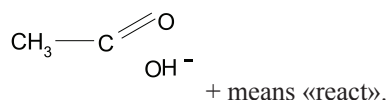
The first feature tests whether the two NPs belong to the same semantic class. This feature is directly connected with clustering, determining the clustering radius and its threshold.



This chemical formula means that acetic acid reacts with ethyl alcohol, resulting in ethyl acetate and water. There are the following acronyms used here:



$\text{C}_2\text{H}_5\text{OH}$ - ethanol



HOH- water

• Positional feature. There is only one positional feature that measures the distance between the two NPs in sentences.

As a rule, coreference has sense only with respect to the specification within one language, e.g. «the language of chemistry», «the Russian language», «the English language», etc. The notions sounding in Russian and in chemistry in a similar way, can have different meanings, and it is possible to reflect it in coreferential NPs made up correctly. Let us discuss the descriptor «energy» («энергия»).

The meanings of NPs in English and Russian languages coincide: электрическая энергия – electric energy, энергичный человек – an energetic man.

The corresponding NPs in chemistry represented in English and Russian are: энергия химической связи – chemical bond energy, энергия химической реакции – the energy of a chemical reaction.

So, the meaning of the same concept in natural languages and in the language of sciences (chemistry, in our case) can be different.

Summing up, we can say that recent years have seen an intensifying interest in NP coreference. This is the problem of determining which NPs refer to the same real-world entity in a document.

As a result of this investigation, various new models and approaches to NP coreference have been worked out. The investigation of new coreference models and new linguistic features can hardly be overestimated.

One of the principle problems nowadays is to establish relations of coherence between ref-

The second feature tests whether one NP is a name alias or acronym of the other. Acronyms were borrowed into linguistics from natural sciences and are used in chemistry to denote all chemical substances. A specialist should understand symbolic formulas, e.g.

erence objects. As chemistry belongs to sciences, coreferential links are in the deep structure of its concepts.

Using coreference in order to overcome communicative failures in teaching chemistry to international students in Russian gives us a possibility to represent the real world from the holistic point of view.

References

1. GOST 7.24-2007 The thesaurus the information retrieval multilingual. Structure, structure and the basic requirements to construction.
2. The thesaurus of organic reactions of chemistry, chemical manufactures // AN USSR. – M.: Science, 1980.
3. Cardie C., Kiri W. Noun phrase coreference as clustering. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora // College Park, MD. – 1999. – P. 82–89.
4. Shallow Ng.V. Semantics for Coreference Resolution // Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI). – 2007. – P. 1689-1694.
5. Nomenclature of organic chemistry (Sect. A. – H., Oxford, Pergamon Press, 1979).
6. Poesio M. Coreference. MATE Dialogue Annotation Guidelines-Deliverable D2.1. – January 2000, 126-182. (2000) (<http://www.ims.uni-stuttgart.de/projekte/mate/mdag>).
6. Soon W.M., Ng H.T., Lim, D.C.Y. A machine learning approach to coreference resolution of noun phrases // Computational Linguistics. – 2001. – №27(4). – P. 521–544.

The work was submitted to the International Scientific Conference «Modern Natural Scientific Education», France (Paris), October, 15–22, 2011, came to the editorial office on 04.08.2011.