COMPUTATIONAL LEXICOGRAPHY:
AN OVERVIEW OF WORDNET AND FRAMENET DICTIONARIES

Elena TROHIN

Catedra Filologie Engleză

Prezentul articol este dedicat lexicografiei computaţionale. Sunt prezentate noţiuni generale din acest domeniu, precum şi informaţii despre două sisteme de dicţionare pe suport electronic de tip nou: WordNet şi FrameNet.

Computational lexicography is the study of making and use of electronic lexicons and dictionaries, encompassing the form, meaning and behaviour of words.

We should mention that computational lexicons and dictionaries include computerized versions of ordinary dictionaries. Lexicons also include any electronic compilations of words, phrases and concepts, such as word lists, glossaries, taxonomies, terminology databases, wordnets and ontologies. A key component of computational lexicons is that they contain at least some additional information associated with the words, phrases or concepts. In general, a lexicon includes a wide array of information associated with entries. An entry in a lexicon is usually the base form of a word, the singular for a noun and the present tense for a verb. More specialized lexicon contains additional types of information. A thesaurus or wordnet contains synonyms, antonyms, or words bearing some other relationship to the entry. Examples of English electronic lexicons are ***WordNet,*** (originally aimed at human users) and ***FrameNet,*** currently the most popular.

The term electronic dictionary can be used to refer to any reference material stored in electronic form that gives information about spelling, meaning, or use of words. Thus a spell-checker in a word-processing program, a device that scans and translates printed words, a glossary for on-line teaching materials, or an electronic version of a respected hard copy dictionary are all electronic dictionaries. [Nesi H., Electronic Dictionaries in Second Language Vocabulary Comprehension and Acquisition, 2000, p.47].

So, we can make the conclusion that electronic dictionaries are collections of structured electronic data that can be accessed with multiple tools, enhanced with wide range of functionalities, and used in various environments.

We can distinguish the following types of electronic dictionaries:

1) dictionary for human users;
2) computer-based dictionaries;
3) machine-readable dictionaries;
4) lexical/term banks;
5) machine dictionaries;
6) lexical databases;
7) artificial intelligence lexicons.

In general, a lexicon includes a wide array of information associated with entries. An entry in a lexicon is usually the base form of a word, the singular for a noun and the present tense for a verb. More specialized lexicon contain additional types of information. A thesaurus or wordnet contains synonyms, antonyms, or words bearing some other relationship to the entry. Examples of English lexicons are ***WordNet,*** (originally aimed at human users) and ***FrameNet,*** currently the most popular.

WordNet is a large lexical database of English, developed under the direction of psychology professor ***George A. Miller*** and Cognitive Science Laboratory of Princeton University**.** Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. Its purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications.

As of 2006, the database contains about 150,000 words organized in over 115,000 synsets for a total of 207,000 word-sense pairs; in compressed form, it is about 12 megabytes in size. WordNet distinguishes between nouns, verbs, adjectives, and adverbs because they follow different grammatical rules. Every synset contains a group of synonymous words or collocations, which are a sequence of words that go together to form a specific meaning, different senses of a word are in different synsets. The meaning of the synsets is further clarified with short defining *glosses* (definitions or example sentences). Most synsets are connected to other synsets via a number of semantic relations. These relations vary based on the type of word, and include:

1) nouns (hypernyms, hyponyms, coordinate terms, holonym, meronym);
2) verbs (hypernym, troponym, entailment, coordinate terms);
3) adjectives (related nouns, similar to, participle of verb);
4) adverbs (root adjectives).

While semantic relations apply to all members of a synset because they share a meaning but are all mutually synonyms, words can also be connected to other words through lexical relations, including antonyms and derivationally related, as well.

WordNet also provides the *polysemy count* of a word: the number of synsets that contain the word. If a word participates in several synsets (i.e. has several senses) then typically some senses are much more common than others. WordNet quantifies this by the *frequency score*: in which several sample texts have all words semantically tagged with the corresponding synset, and then a count provided indicating how often a word appears in a specific sense.

The morphology functions of the software distributed with the database try to deduce the lemma or root form of a word from the user's input, only the root form is stored in the database unless it has irregular inflected forms. The goal of WordNet was to develop a system that would be consistent with the knowledge acquired over the years about how human beings process language. WordNet can be interpreted and used as a lexical ontology. We should mention that WordNet does not include information about etymology, pronunciation and the forms of irregular verbs and contains only limited information about usage. The actual lexicographical and semantic information is maintained in *lexicographer files*, which are then processed by a tool called *grind* to produce the distributed database. Though WordNet contains a sufficiently wide range of common words, it does not cover special domain vocabulary. WordNet is also freely and publicly available for download. WordNet is considered to be the most important resource available to researchers in computational linguistics, text analysis, and many related areas.

FrameNet is a project housed at the International Computer Science Institute in Berkeley, California. FrameNet project is creating an on-line lexical resource for English, based on frame semantics and supported by corpus evidence. The aim is to document the range of semantic and syntactic combinatory possibilities (valences) of each word in each of its senses, through computer-assisted annotation of example sentences and automatic tabulation and display of the annotation results. The major product of this work, the FrameNet lexical database, which currently contains more than 11,600 lexical units (defined below), more than 6,800 of which are fully annotated, in more than 960 semantic frames, exemplified in more than 150,000 annotated sentences. It has gone through five releases, and is now in use by hundreds of researchers, teachers, and students around the world.

FrameNet data is available online as browsable reports, a clickable visualization, and a searchable database. You can also download the data in XML format. In addition to its lexicographic work, FrameNet has begun to annotate some continuous texts, as a demonstration of how frame semantics can contribute to text understanding. It is important to mention that FrameNet lexical units come with the definitions from the Concise Oxford Dictionary, tenth edition or a definition written by a FrameNet staff member. Unlike commercial dictionaries FrameNet provides multiple annotated examples of each sense of a word .The set of examples illustrates all of the combinatorial possibilities of the lexical unit. The examples are attestations taken from naturalistic corpora rather than constructed by a linguist or a lexicographer. The main FrameNet corpus is the 100-milion-word British National Corpus, which is both large and balanced across genres (editorials, text-books, advertisements, novels, etc.) But we should mention that it lacks many specifically American expressions.

Its analysis of the English lexicon proceeds frame by frame rather than by lemma, whereas traditional dictionary making proceeds word by word through the alphabet. So, we can make the conclusion that if a traditional lexicography measures progress in words completed, FrameNet measures progress in frames completed. Each lexical unit is linked to a semantic frame, and to the other words which evoke that frame. We can notice that this makes the FrameNet database similar to a thesaurus, grouping together semantic similar words.

Bibliography:

1. Баранов А.Н. Компьютерная лексикография // Баранов А.Н. Введение в прикладную лингвистику. - Москва: УРСС, 2001, с.81-88.
2. Беляева Л.Н., Герд А.С., Убин И.И. Автоматизация в лексикографии // Прикладное языкознание: Учебник / Под ред. А.С. Герда. - Санкт-Петербург: Изд-во СПб-ского ун-та, 1996, с.318-333.
3. Леонтьева Н.Н. К теории автоматического понимания естественных текстов. Часть 2. Семантические словари: состав, структура, методика создания. - Москва: Изд-во Моск. ун-та, 2000. - 40 с.
4. Перцов Н.В.; Старостин С.А. О лексикографической справочной информационной системе ЛЕКСИС по русскому языку // Труды Международного семинара "Диалог '95" по компьютерной лингвистике и ее приложениям" = "Dialogue '95. Computational linguistics and its applications" international workshop, Казань, 31 мая-4 июня 1995 г. - Казань, 1995, с.247-249. Рез. англ.
5. http://framenet.icsi.berkeley.edu/
6. http://wordnet.princeton.edu/

Prezentat la 23.12.2009