Modena Lexi-Term

Dictionaries: resources and perspectives
Dictionnaires: ressources et perspectives
Workshop/Atelier de recherche
Modena, 5 November 2009/Modène, le 5 novembre 2009

ABSTRACTS

Geoffrey Williams
(Université de Bretagne-Sud)

Going Natural: Building Organic Dictionaries

Whether your dictionary be aimed at so-called general language or what, for want of a better term, the so-special language dictionaries, one major problem is that same, namely what words to include and what to leave out.
Traditional terminology processing solved the problem by being top-down and imposing a fixed onomasiological framework into which, largely nominal 'terms' were incorporated. The technological explosion coupled with a relation of the needs to take into account real-life situations has meant that this is no longer tenable. The necessity for a change in outlook has become all the more pressing with the rise of corpus linguistics and NLP, both of which were initially slow to penetrate the prescriptive world of terminology.
In lexicography two situations can be found; inclusion of specialised usage in general language dictionaries and the construction of discipline specific dictionaries. The former is notably non-systematic in its approach, even when claims are made for essential base vocabulary lists in learner's dictionaries. The second ends to be written by specialists and are both encyclopaedic in nature and biassed towards the particular approach of their author. Corpora are not used.
The aim here is to concentrate on scientific dictionaries aimed at encoding rather than decoding and their creation from corpora.
In any dictionary project, the first problem is the audience. In this case we are essentially aiming at scientists, both confirmed scientists and young researchers publishing in English. The hypothesis is that these have a good knowledge of their terminology, but lack genre-specific writing skills. The aim thus is to essentially produce a pattern dictionary illustrating usage in specific environments.
The second stage is to define the area to be covered. Hierarchies of disciplines can be one approach, but in real life most studies are multidisciplinary so a thematic approach to corpus building and analysis will need to be taken into account.
The third problem is what words to include, and what to leave out. This brings us to the notion of organic dictionaries based on collocational networks and the construction of natural ontologies. After a discussion of the problems of corpus building we shall look at two experimental dictionaries; the Parasitic Plant Dictionary and the SCIENTEXT Verb Dictionary. The former is thematically oriented and entirely built using collocational frameworks derived from key words. The second is based on a general corpus of scientific research articles and takes as its starting point the 100 most frequent verbs. The latter is based on the open-source BMC corpus which will be available shortly on line. This corpus is part of speech tagged and lemmatised and prepared for use with XAIRA.

*

Silvia Cacchiani
(Università di Modena e Reggio Emilia)

On the genuine purpose of the dictionary
Some reflections on genuine purpose, cognition- and knowledge-oriented functions of English law dictionaries

Commercial (paper) dictionaries are *utility products* (Wiegand 1998). They are designed and compiled in order to meet the needs of specific target users in a specific situation of use. This is known as the *genuine purpose* of the dictionary, which also covers subject field area and cognition- and knowledge-oriented functions of the dictionary (Bergenholz/Nielsen 2006).
It is against this background that we contrast three English law dictionaries (*The Longman Dictionary of Law*, *The Law Student's Dictionary*, *Oxford Dictionary of Law*) with an eye to their pedagogical dimension (if any) and the needs of Italian law students enrolled in Legal English courses. Specifically, the emphasis lies on

meaning representation (Wiegand 1992ff.) and inclusion of encyclopaedic and non-encyclopaedic information in dictionary entries.

*

Serena Sorrentino
(Università di Modena e Reggio Emilia)

Semi-automatic schema labels normalization for improving schema matching

Schema matching is the problem of finding relationships among concepts across heterogeneous data sources (heterogeneous in format and in structure). Starting from the “hidden meaning” associated to schema labels (i.e. class/attribute names) it is possible to discover relationships among the elements of different schemata. Lexical annotation (i.e. annotation w.r.t. a thesaurus/lexical resource) helps in associating a "meaning" to schema labels. However, accuracy of semi-automatic lexical annotation methods on real-world schemata suffers from the abundance of non-dictionary words such as compound nouns and word abbreviations. In this work, we address this problem by proposing a method to perform schema labels normalization which increases the number of comparable labels. Unlike other
solutions, the method automatically expands abbreviations and annotates compound terms, without a manual effort. We prove that our normalization method helps in the identification of similarities among schema elements of different data sources, thus improving schema matching accuracy.

*

Laura Po
(Università di Modena e Reggio Emilia)

Lexical Knowledge Extraction: an effective approach to schema and ontology matching

The aim of this talk is to examine what role Lexical Knowledge Extraction plays in data integration as well as ontology engineering. Data integration is the problem of combining data residing at distributed heterogeneous sources, and providing the user with a unified view of these data; a common and important scenario in data integration are structured or semi-structure data sources described by a schema. Ontology engineering is a subfield of knowledge engineering that studies the methodologies for building and maintaining ontologies. Ontology engineering offers a direction towards solving the interoperability problems brought about by semantic obstacles, such as the obstacles related to the definitions of business terms and software classes. In these contexts where users are confronted with heterogeneous information it is crucial the support of matching techniques. Matching techniques aim at finding correspondences between semantically related entities of different schemata/ontologies. Several matching techniques have been proposed in the literature based on different approaches, often derived from other fields, such as text similarity, graph comparison and machine learning. We propose a matching technique based on Lexical Knowledge Extraction: first, an Automatic Lexical Annotation of schemata/ontologies is performed, then lexical relationships are extracted based on such annotations. Lexical Annotation is a piece of information added in a document (book, online record, video, or other data), that refers to a semantic resource such as WordNet. Each annotation has the property to own one or more lexical descriptions. Lexical annotation is performed by the Probabilistic Word Sense Disambiguation (PWSD) method that combines several disambiguation algorithms. Our hypothesis is that performing lexical annotation of elements (e.g. classes and properties/attributes) of schemata/ontologies makes the system able to automatically extract the lexical knowledge that is implicit in a schema/ontology and then to derive lexical relationships between the elements of a schema/ontology or among elements of different schemata/ontologies. The effectiveness of the method presented has been proven within the data integration system MOMIS (Beneventano D. *et al.* 2003).

*

Luciana T. Soliman
(Università di Modena e Reggio Emilia)

Ce qu'on ne doit pas dire: le texte et le terminaire

La fonction prescriptive de la terminologie vise à la production d'une communication claire et efficace. Ses orientations se heurtent pourtant à l'usage effectif où une synonymie « physiologique » survit aux opérations de standardisation. Il faut tenir compte non seulement des facteurs sociolinguistiques (usage établi, milieu d'implantation, besoins des usagers), mais aussi des facteurs psycholinguistiques (motivation, habitudes du locuteur/scripteur, résistance naturelle aux changements). Dans les textes concernant la micro-informatique,

par exemple, les formations endogènes et exogènes cohabitent dans un flou terminologique qui s'oppose au dirigisme linguistique. Certains produits terminologiques à vocation strictement prescriptive s'alignent sur les recommandations des commissions de néologie et de terminologie, mais s'éloignent ainsi de la réalité linguistique; d'autres signalent ces remarques, mais optent pour une harmonisation moins contraignante des emplois, l'aspect fréquentiel jouant un rôle capital.

***