Bilingual Dictionaries and Lexical Gaps

Maarten Janssen
Universteit Utrecht

1 Introduction

With the ongoing cooperation between countries with different languages (such as in the forthcoming expansion of the European Union) there is a growing need for bilingual dictionaries. Given the explosive growth of the number of dictionaries required with an increasing number of langugages (for 16 languages one already needs 256 dictionaries), the need for a more principled and automated approach to the generation of bilingual dictionaries is not hard to establish.

The main approaches in this field can be divided into two groups: on the one hand the theory based approaches, such as EuroWordNet (Vossen, 1998) and the Hub-and-Spoke Model (Beeken *et al.*, 1998). And on the other hand the corpus based appoaches such as COBUILD (Sinclair, 1987) and PAROLE. This paper clearly falls within the first tradition.

If every word in a language would have a (near) perfect translation in every other language, linking languages would be relatively easy. The real problems start when a language lacks a translation for a word, which is often referred to as a lexical gap. As was shown in a study by the MultiWordNet project group (Bentivogli & Pianto, 2000), these lexical gaps do not form a small, isolated group: for the English-Italian Collins bilingual dictionary, 7.8% of the words is a lexical gap.

To take an oversimplified example, consider the English word *colt*. There is no clear translation for *colt* in French: French only has the word *poulain*,which means *foal*. Such lexical gaps necessarily lead to a tension: for productive use, one needs a word that can be used in text as a translation for *colt*. Which word this is, can be established with corpus analysis, which will show that *poulain* is the translation used for *colt*. So corpus methods, with all their problems, could work for production-oriented bilingual dictionaries, such as an English-French dictionary written for an English speaking audience.

On the other hand, for a perception-oriented dictionary this does not suffice: the French speaker reading an English text does not only want to know a *colt* is a *poulain*, but that it is specifically a *male* foal. An English-French dictionary for a French audience should give the translationally equivalent description: *poulain male*. And since this description is never used as a translation, there seems little change corpus analysis of any kind will render this information.

In my PhD project, this information should be available, by drawing it from its most natural source: the English monolingual dictionary, in which a *colt* is defined as being a *male foal*. The project aims at aligning languages by means of their monolingual definitions. A logical framework, called Formal Concept Analysis, assures that *colt* then gets properly related to *poulain male*.

References

Beeken, Jeannine; Heid, Ulrich; Laureys, Godelieve; Martin, Willy & Schuurman, Ineke. 1998. *On the Construction of Bilingual Dictionaries: feasibility study carried out by order of the European Commission DG XIII*. Technical Report. Stuttgart.

Bentivogli, Luisa & Pianto, Emanuele. 2000. Looking for Lexical Gaps. *In: Proceedings of the Ninth Euralex International Congress*.

Sinclair, J.M. 1987. *Looking Up: an Account of the COBUILD Project in Lexical Computing*. Collins.

Vossen, Piek. 1998. Introduction to EuroWordNet. *Computers and the Humanities*, vol. 32:73 – 89. 1998.