Annotated corpora for assistance with English-Polish translation.

Barbara Lewandowska-Tomaszczyk, University of Lodz, Poland,
Michael P. Oakes, University of Sunderland, England,
Paul Rayson, University of Lancaster, England.

Alignment of a large bilingual corpus of original material and an acceptable translation facilitates a number of automated and partially automated approaches to translation. (Kay and Roescheisen 1993, p 121). The approach to the automatic alignment of Polish and English texts taken in this paper is that of Gale and Church (1993). Once the texts have been aligned, they can then be displayed to the translator as required using Scott's (1996) "WordSmith" concordancing tool. Using WordSmith, sentences and their translations can be retrieved and shown to the translator if they contain specified words, phrases or word fragments. The power of the bilingual concordancing tool can be enhanced by using annotated corpora for alignment.

The two types of annotation we have employed are a) Part of Speech tagging, provided by the CLAWS tagger (Garside and Smith, 1997), and b) semantic tagging, provided by the ACASD suite of computer programs (Thomas & Wilson, 1996). With part of speech tagging, every word in the corpus is automatically assigned its most likely grammatical class, e.g. the notation "book_NN1" shows that a particular instance of the word "book" is most probably a singular common noun. If each word in the aligned text has its own part of speech category, we can use the search term "book" to retrieve every aligned region containing the word "book", irrespective of whether it occurs as a noun, verb, or any other grammatical category. On the other hand, the notation "book_NN1" will only retrieve aligned regions in which the word "book" occurs as a singular common noun, enabling us to see the various translations in Polish of the English word "book" when it occurs as a singular common noun.

Semantic tagging involves assigning each word in the corpus with a label showing the semantic category (akin to a thesaurus category such as "colour", "power" or "education") to which each word belongs. For example, the notation "clinical_B2" would retrieve only aligned regions containing "clinical" as a medical term, while "clinical_E1-" would retrieve only those regions containing the word "clinical" where it means "lacking emotion". The notation "clinical" allows the retrieval of aligned regions containing the word "clinical" irrespective of its meaning, "_B2" will allow the retrieval of aligned regions containing any medical term, and "_E1-" will allow us to view all the synonyms in the English corpus meaning "without emotion" alongside their Polish equivalents.

At present only the English text has been fully annotated, since a Polish semantic tagger is currently unavailable. We will discuss our approach whereby a machine readable Polish-English dictionary and an alignment of semantically tagged English text with the equivalent unannotated Polish text might be used to semantically tag certain words in the Polish part of the text.

References

Gale W A, Church K W 1993 A program for aligning sentences in bilingual corpora, *Computational Linguistics* 19(1): 75-102.

Garside R, and Smith N 1997 A hybrid grammatical tagger: CLAWS4, in Garside R, Leech G, and McEnery A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pp. 102-121.

Kay M, Roescheisen M 1993 Text-translation alignment, *Computational Linguistics* 19(1):121-142..

Scott M 1996 *WordSmith Tools Manual*, Oxford University Press.

Thomas J, Wilson A 1996 Methodologies for studying doctor-patient interaction, in Thomas J, Short M (eds) *Using corpora for language research*, London & New York, Longman.