li179 Linguistic Insights
Studies in Language and Communication

Ana Díaz-Negrillo &
Francisco Javier Díaz-Pérez (eds)

Specialisation and Variation in Language Corpora

Peter Lang

ANA DÍAZ-NEGRILLO / FRANCISCO JAVIER DÍAZ-PÉREZ

Trends in corpus specialisation

1. Introduction

Computerised corpus linguistics set off around the 1960s with the compilation and exploitation of the first reference corpus of the English language. Over 50 years later, reference corpora are probably the largest in size and most consolidated corpus types. They are also perhaps the corpus type that reaches the largest number of users, as they are used by both specialists and non-specialists in linguistics. This is so much so that they are increasingly regarded as another reference tool of language use.

While English reference corpora have become consolidated, corpus linguistics and language corpora have also acquired a remarkable degree of expansion and specialisation. The continuous growth of corpus linguistics has been fostered by the interest of users in language-related areas, who have realized the powerful tool corpora can be in their disciplines. Nowadays, a large number of language corpora of an extensive variety of languages exist. Indeed, national corpora of major languages are available, as well as corpora of languages spoken by smaller communities. Corpora currently also cover a range of registers, text types and subject fields. Actually, the increasing specialisation in corpus linguistics has made it possible to investigate a variety of linguistic aspects for a range of applications in a variety of linguistic areas, for example, language teaching, second language acquisition, translation, terminology, stylistics, discourse analysis, etc.

Progress in the design and implementation of data processing tools has also played a crucial role in the development of corpus linguistics. Originally, most of these tools were designed for written,

native, non-specialised corpus data, some of which were later transferred to specialised corpora. However, in order to suit the particular features represented by the language or text type in question, specific tools were necessary in order to cater for specialised corpus data and the methodological approaches required in a variety of specific domains.

This volume is further evidence of the extraordinary expansion that corpus linguistics and language corpora have gone through over the past years. It focuses on emerging corpus types, corpus techniques and corpus-based linguistic studies in areas that can now be researched as a result of the recent development of corpus linguistics, and which were presented at the *4th International Conference on Corpus Linguistics* '*Language, Corpora and Applications: diversity and change*' (22-24 March 2012, University of Jaén, Spain). In so doing, this volume also intends to support work in corpus linguistics, which may lead to the initiation, development or consolidation of new approaches to the design, processing and analysis of language corpora.

In order to give evidence of the expansion of language corpora, the volume covers small and specific corpora, both as to the languages represented in the corpora (Basque, Greek, Catalan, Hungarian), and the type of corpora they represent (learner corpora, translation corpora, correspondence corpora and technical −medical− corpora). Specifically, it comprises papers on technical aspects of corpus data processing (section 1), on corpus-based linguistic research (section 2), and on emerging corpora (section 3), all of which will be discussed, in this order, in the rest of this chapter.

2. Corpus-data processing

The first section in the volume deals with procedural issues that are central in corpus linguistics: corpus annotation in written and oral corpora, automatic identification and extraction of linguistic items,

and corpus multimodality. The section is mainly occupied with types of corpora associated with areas of applied linguistics (learner and translation corpora) and which, due to their nature, require special corpus analysis techniques.

TANTOS/PAPADOPOULOU and HERMENT ET AL. give evidence of the development of learner corpora in recent years. Learner corpora began to be compiled around the 90s as collections of written material of non-native language to be used for pedagogical and SLA research purposes (Granger 2002). In terms of corpus annotation, and due to their language-specific features, learner corpora have largely relied on manual annotation, specifically error and interlanguage annotation, and on tools which were designed for native corpus data, like POS tagging (Granger et al. 2009) (for an overview, see Díaz-Negrillo/Thompson 2013). In recent years, however, the identification of learner-specific features in corpus data has started to become at least partially automatized, and the annotation procedures have also become more sophisticated.

In this volume, TANTOS/PAPADOPOULUS look at error annotation of a corpus of learner Greek: the Greek Learner Corpus (GLC). Their work stands among the first initiatives to compile a corpus of Greek learner language (cf. also Tzimokas 2010) and, in particular, it stands out for the formalisation of its error annotation in a multi-layered fashion. The tagset has been designed following the hierarchical structure of error categories originally proposed by Dagneaux et al. (1996) and is implemented in the corpus using UAM corpus tool (O'Donnell 2008), a software which stores multi-layered corpus annotations.[1] This freeware is used nowadays for a variety of manual annotation types. Some outstanding features are that it requires no expertise in programming on the part of the user and that it is rather versatile containing also a tool for statistic analysis. Finally, the paper explains the annotation standard used for the corpus. While the TEI[2] has been widely used as a format standardisation purposes in language corpora (cf., however, also TUSNELDA[3]), the GLC uses the stand-off

1 Downloadable from <http://www.wagsoft.com/CorpusTool/>
2 <http://www.tei-c.org/index.xml>
3 <http://www.sfb441.uni-tuebingen.de/c1/tusnelda-guidelines.html>