Legal canvas for a patchwork of multilingual quotations: the case of CoMParS

Piotr Bański, Paweł Kamocki and Beata Trawiński (IDS Mannheim, Germany)

1. Introduction

CoMParS is a resource under construction in the context of the long-term project German Grammar in European Comparison (GDE) at the IDS Mannheim. The principal goal of GDE is to create a novel contrastive grammar of German against the background of other European languages. Alongside German, which is the central focus, the core languages for comparison are English, French, Hungarian and Polish, representing different typological classes.

Unlike traditional contrastive grammars available for German, which usually cover language pairs and are based on formal grammatical categories, the new GDE grammar is developed in the spirit of functionalist typology. This implies that, instead of formal criteria, cognitively motivated functional domains in terms of Givón (1984) are used as *tertia comparationis*.

The purpose of CoMParS is to document the empirical basis of the theoretical assumptions of GDE-V and to illustrate the otherwise rather abstract content of grammar books by as many as possible naturally occurring and adequately presented multilingual examples, including information on their use in specific contexts and registers. These examples come from existing parallel corpora, and our presentation will focus on the legal aspects and consequences of this choice of language data.

2. Motivation and general assumptions

Corpus-based data are particularly precious because they present real context- and register-related language usage. At the same time, however, they are often not suitable for exemplification purposes because they are too complex and contain big portions of irrelevant material. The existing parallel corpora can already serve as a solid data source for contrastive research, but due to missing or sparse linguistic annotation, they do not reach their full potential. This is particularly true of multilingual parallel corpora: only two of them, namely InterCorp with 38 languages and 1.5 billion tokens (Čermák and Rosen 2012) and ParaSol with 31 languages and 27 million tokens (Waldenfels 2006) are lemmatized and grammatically annotated.

However, linguistic annotation in InterCorp and ParaSol identifies morphosyntactic properties alone; functional-semantic annotation is not available. Moreover, grammatical information available in these corpora is language-specific and thus not directly comparable across languages. As a result, only language-specific form-based queries can be performed, as opposed to meaning- or function-related queries such as, for example, "How is REFERENCE / QUANTIFICATION / REFLEXIVITY / POSSESSION / EXPERIENCE etc. expressed in languages L1...L*n*"? However, precisely this kind of research question is substantial for crosslinguistic studies conducted within GDE. Obtaining parallel sequences appropriately exemplifying specific communicative functions across different languages from the available corpora using form-based queries is a laborious procedure. Moreover, the extracted data are usually more complex than needed to illustrate the point at hand, and for this reason, they are not suited to be directly utilized as examples.

It is against this background that CoMParS is being constructed, as a small multilingual database of parallel sequences annotated with semantic-functional information and designed especially for the purposes of data-driven contrastive research, in particular contrastive grammar writing, with a view to language-didactic applications.

The general idea behind our approach to building CoMParS is to extract data out of the existing parallel corpora using the usual corpus query tools. The extracted data are then carefully examined by grammar writers, checked for quality of translation and accuracy of functional equivalence, as well as for relevance and applicability as examples in a (multilingual) contrastive grammar. We may call this a process of "refinement".

Next, those parts of the selected aligned data are identified that are the smallest necessary for the appropriate exemplification purposes. Exactly those parts get additionally annotated with semantic-functional information. At the same time, it may happen that missing members of multilingual n-tuples are constructed (and clearly marked as such). This can be thought of as "enrichment".

The extracted n-tuples also include information present in the original metadata and a link to the original resource from which they are cited.

3. Legal issues surrounding the construction of CoMParS

Intellectual Property Rights (IPR), in particular copyright and the *sui generis* database right have been identified as one of the major obstacles for the creation of language corpora. CoMParS is also heavily affected by these issues, as shown below.

Firstly, copyright needs to be addressed, because a large part of the material used in CoMParS is still under copyright protection (which expires 70 years after the death of the author). Indeed, in order to reproduce a copyright-protected work (including excerpts therefrom) and communicate it to the public, one needs an authorisation from the rightholder, unless the use enters within the scope of a statutory exception. CoMParS enters within the scope of the quotation exception under German law, and possibly other national laws of the EU Member States. International (art. 10.1 of the Berne Convention) and EU law (art. 5.3(d) of the Copyright Directive) require that in order to be lawful, quotation has to be justified by its purpose. In some jurisdictions, only excerpts of works can be quoted (e.g. in France), while others allow whole works to be quoted under specific circumstances (e.g. in Germany, where s. 51(1) UrhG allows 'long scientific quotations'). Furthermore, most jurisdictions (including Germany and France) require that the citation is included in an independent work (i.e., that the quoting work meets the requirements for copyright protection, ie. is its author's own intellectual creation, even if the quotations were removed), which means that a mere compilation of citations, without any original contribution, is not allowed. While this would arguably exclude many language corpora from the scope of this exception, this is not the case of CoMParS, which, while based on data extracted from existing corpora, is also enriched/extended by additional annotations, functional alignments, and, in some cases, also additional content, resulting in work derived from, and not merely copying, the original sources.

Moreover, the Court of Justice of the European Union recently ruled (C-145/10, *Painer*), that the Copyright Directive does not require the quoting work to meet the criteria for copyright protection. Member States are thus allowed to abandon this requirement (which is what Slovakia did in 2015), but are not obliged to do so. Even before the *Painer* case, courts in jurisdictions like France or Germany on occasions adopted an extensive interpretation of the quotation exception. This is what happened in the *Germania 3* ruling in Germany (in which the Federal Constitutional Court ruled that copyright rules (and exceptions) should be interpreted in a way as not to inhibit basic freedoms, such as freedom of artistic expression or freedom of research) and the *Microfor* case in France (in which the Court of Cassation ruled that a work made for informational purposes — and (electronic) collections of linguistic data are likely to qualify in this category — can quote other works even if it does not itself meet the criteria for copyright protection).

Secondly, copyright may also protect compilations (such as language corpora) if the selection and arrangement of their contents is original (ie. is its author's own intellectual creation). In such cases, however, copyright only protects the 'envelope' (selection and arrangement), but not the contents. While CoMParS does reproduce data from other corpora, it does not copy their original selection and arrangement.

Finally, language corpora are likely to qualify as databases and therefore they may be protected by a *sui generis* database right if there was a substantial investment involved in their creation. The holder of the *sui generis* database right (i.e. the investor – in case of language corpora usually a research institution or a research funding agency) can prohibit extraction and re-utilisation of substantial parts of the database, regardless of its originality. While it is extremely unclear what constitutes a substantial part of a database, CoMParS only extracts and re-utilizes quantitatively small parts of the above mentioned corpora; therefore, the *sui generis* database right is not infringed.

4. Conclusion

The present contribution sketches the legal aspects of a decision to create an electronic language resource on the basis of quotations from other such resources, enriched by original content and structured according to semantic-functional criteria, in the context of EU law, and specifically German and French norms. We see this work as explaining the legal assumptions of CoMParS and, on a broader scale, as contributing to the search for legal solutions concerning the creation of complex language resources. We hope to have drawn the community's attention to the potential use of the quotation exception (rather than merely the research exception) for the purpose of language resource construction.

References

Čermák, F. & A. Rosen (2012). The case of InterCorp, a multilingual parallel corpus. In: *International Journal of Corpus Linguistics*, 17(3), pp. 411-427.

Givón, T. (1984). *Syntax. A functional-typological introduction*. Amsterdam, Philadelphia: J. Benjamins.

v. Waldenfels, R. (2006). Compiling a parallel corpus of slavic languages. Text strategies, tools and the question of lemmatization in alignment. In: Brehmer, B., Zdanova, V., Zimny, R. (Eds). *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9*. München, pp. 123-138.

Legislative acts

Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979).

Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.

Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.

Gesetz über Urheberrecht und verwandte Schutzrechte (German copyright act, UrhG).

Zákon č. 185/2015 Z. z. Autorský zákon (Slovak copyright act).

Court cases

Cour de cassation, Assemblé plénière, 30 October 1987, no. 86-11.918 (*Microfor*).

Bundesverfassungsgericht, 29 June 2000, 1 BvR 825/98 (*Germania 3*).

Court of Justice of the European Union, Third chamber, 1 December 2011, C-145/10 (*Painer*).