Towards a Regionally Balanced Corpus of Standard German

Freitag,
06.03.2020
12:15–12:45
VMP5 2101/2105

Andreas Nolda, Adrien Barbaresi, Alexander Geyken
Berlin-Brandenburgische Akademie der Wissenschaften
andreas.nolda@bbaw.de, barbaresi@bbaw.de, geyken@bbaw.de

In this talk we present the design of a regionally balanced corpus of present-day written Standard German. It will be compiled from local news sections of recent issues of German, Austrian, and Swiss newspapers. The journals are classified with respect to a set of 8 regions (diatopic areas), also used in the *Variantenwörterbuch des Deutschen* (VWB²; Ammon *et al.* 2016) and the *Variantengrammatik des Standarddeutschen* (VG; http://mediawiki.idsmannheim.de/VarGra/). We aim at a total of 40 journals, i.e. 5 journals per area. Corpus queries can refer to this areal metadata in a faceted search; in addition, linguistic annotation layers for lemmatization and part-of-speech tagging are provided. The user interface will present the query results in various forms, including, *inter alia*, tabular and cartographic representations of the regional distribution in terms of absolute or relative frequen-cies. The corpus will be made publicly accessible for registered users studying regional variation. A major target group are lexicographers interested in improving the empirical accurateness of diatopic information in German dictionaries.

A corpus of this kind has been described as a desideratum by Bickel, Hofer, and Suter (2015: 544) in an article on the design of the VWB². For lack of such a corpus, the VWB² project used the commercial "wiso" newspaper database by GBI-Genios Deutsche Wirtschaftsdatenbank GmbH, which, however, is missing linguistic annotations. The VG project, in turn, compiled a web corpus of local news sections of newspapers (Datenerhebung 2018); this corpus, albeit regional balanced and linguistically annotated, has not been made available to the general linguistic community due to licensing issues.

References: Ammon, Ulrich et al. (2016). Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz und Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen. 2nd ed. Berlin: Walter de Gruyter. Bickel, Hans, Lorenz Hofer & Sandra Suter (2015). Variantenwörterbuch des Deutschen (VWB) – NEU: Dynamik der deutschen Standardvariation aus lexikografischer Sicht. In Regionale Variation des Deutschen: Projekte und Perspektiven, ed. by Roland Kehrein, Alfred Lameli, & Stefan Rabanus. Berlin: de Gruyter, 541–562. Datenerhebung (2018). In Variantengrammatik des Standard-deutschen: Ein Online-Nachschlagewerk. http://mediawiki.ids-mannheim.de/VarGra/index.php/Datenerhebung [26 July 2019].