56. Jahrestagung des Leibniz-Instituts für Deutsche Sprache vom 10. bis 12. März 2020

DEUTSCH IN EUROPA

Sprachpolitisch, grammatisch, methodisch

Mittwoch, 11. März 2020, 11:30 Uhr

VON MONOLINGUALEN KORPORA ÜBER PARALLEL- UND VERGLEICHSKORPORA ZUM EUROPÄISCHEN REFERENZKORPUS EURECO

Marc Kupietz und Beata Trawiński (IDS Mannheim)

Heutzutage stehen der korpusorientieren (vergleichenden) Sprachwissenschaft verschiede Korpustypen als Datenquelle zur Verfügung: monolinguale (National)korpora, Parallelkorpora und Vergleichskorpora. Monolinguale Korpora zeichnet sehr hohe und kontrollierte Sprachqualität aus, da sie u.a. fast ausschließlich Originaltexte enthalten. Auf der inhaltlicher Ebene und bezüglich der Größe sind monolinguale Korpora aufgrund unterschiedlicher Zusammensetzungen jedoch nicht wirklich vergleichbar, was ein ernsthaftes Problem für den Sprachvergleich darstellt. Parallelkorpora enthalten hingegen Texte in mehreren Sprachen, denen eine inhaltlich-funktionale Äquivalenz zugrunde liegt und die somit als ideales tertium comparationis verwendet werden können. Andererseits sind übersetzte Texte mit Problemen wie shining through und Normalisierung (Teich 2003) oder Vereinfachung und Explikation (Baker 1996) verbunden. Bei hoher Vergleichbarkeit zeigen Parallelkorpora also eine niedrigere Sprachqualität. Es besteht also ein klarer Bedarf an multilingualen Korpora, die einerseits hohe Vergleichbarkeit auf der inhaltlichen Ebene und hinsichtlich der Größe und andererseits eine originalsprachliche Qualität sicherstellen. Eine interessante Alternative bieten Vergleichskorpora an. Bis jetzt stehen jedoch kaum Vergleichskorpora zur Verfügung, die die obengenannten Kriterien in zufriedenstellendem Maße erfüllen.

Die Idee der 2013 gegründeten EuReCo-Initiative (Kupietz et al. 2018) ist, nicht neue Korpora dieser Art aufzubauen, sondern bereits existierende Referenz- und Nationalkorpora virtuell zu Paaren vergleichbarer Korpora zusammenzuschließen. Das heißt, dass die jeweiligen Korpora an ihren Standorten verbleiben und über eine gemeinsame Softwareinfrastruktur, zurzeit die Korpusanalyseplattform KorAP (Bański et al. 2013), vernetzt und über eine einheitliche Schnittstelle zur Verfügung gestellt werden. Die Konstruktion vergleichbarer Korpora erfolgt dann anhand von Textmetadaten so, dass idealerweise der Nutzer selbst dynamisch virtuelle vergleichbare Subkorpora definieren kann – etwa durch Kommandos wie „bilde das größtmögliche Korpuspaar mit identischer Zusammensetzung bezüglich Thema, Texttyp und Veröffentlichungsjahr". Eine solche dynamische Definierbarkeit, mit Möglichkeit zur persistenten Speicherung, ist wichtig, da aufgrund der zusätzlichen Anforderung der „Vergleichbarkeit" an die Stichprobenpaare das Risiko durch Korpuskomposition bedingter Artefakte besonders hoch ist. Entsprechend sollte der Konstruktionsprozess grundsätzlich iterativ anlegbar sein (Kupietz 2015), sodass Korpuskompositionen korrigiert werden können und auch die Stabilität quantitativer Befunde bezüglich unterschiedlich definierter vergleichbarer Korpora überprüft werden kann. Mit DruKoLA (Cosma et al. 2016) und DeutUng (Kupietz et al. 2018) liefen bzw. laufen bereits zwei Projekte im EuReCo-Kontext und erste vergleichbare deutsch-rumänische und deutsch-ungarische Korpora werden über KorAP zu sprachvergleichenden Studien genutzt.

Dieser Beitrag beschreibt die Motivation und Ziele hinter der EuReCo-Initiative, präsentiert die Ergebnisse von DruKoLa und DeutUng und skizziert neue Perspektiven für germanistische und vergleichende Korpuslinguistik, insbesondere im europäischen Kontext.

Literatur:

Baker, Mona (1996): Corpus-based Translation Studies: The challenges that lie ahead. In H. Somers (Hg.): Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager. Amsterdam: John Benjamins, 175-187.

Bański, Piotr/Bingel, Joachim/Diewald, Nils/Frick, Elena/Hanl, Michael/Kupietz, Marc/Pęzik, Piotr/Schnober, Carsten/Witt, Andreas (2013): KorAP: the new corpus analysis platform at IDS Mannheim. In Vetulani, Zygmunt/Uszkoreit, Hans (Hrsg.): Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conference, 586-587. Poznań: Fundacja Uniwersytetu im. A. Mickiewicza, 2013.

Cosma, Ruxandra/Cristea, Dan/Kupietz, Marc/Tufiş, Dan/Witt, Andreas (2016): DRuKoLA – Towards Contrastive German-Romanian Research based on Comparable Corpora. In Bański, Piotr/Barbaresi, Adrien/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Kupietz, Marc/Lüngen, Harald/Witt, Andreas: 4th Workshop on Challenges in the Management of Large Corpora. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slowenien. Paris: European Language Resources Association (ELRA), 2016, 28-32.

Kupietz, Marc (2015): Constructing a Corpus. In Durkin, Philip: The Oxford Handbook of Lexicography. (= Oxford handbooks in linguistics). Oxford: Oxford University Press, 62-75.

Kupietz, Marc/Cosma, Ruxandra/Cristea, Dan/Diewald, Nils/Trawiński, Beata/Tufiş, Dan/Váradi, Tamás/Wöllstein, Angelika (2018): Recent developments in the European Reference Corpus (EuReCo). In Granger, Sylviane/Lefer, Marie-Aude/Aguiar de Souza Penha Marion, Laura (Hrsg.): Using Corpora in Contrastive and Translation Studies Conference (5th edition). Book of Abstracts. Louvain-la-Neuve: CECL, 2018, 101-103.

Teich, Elke (2003): Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts. Berlin: Mouton de Gruyter.