

# Korpora aus dem Netz

## Die Erstellung eines Fachkorpus aus Webseiten

Magisterarbeit von Stefan Träger  
Institut für deutsche Sprache und Linguistik

Im Kontext der Linguistik ist ein Korpus „... a collection of naturally-occurring language text, chosen to characterize a state or variety of language.“<sup>1</sup>, also eine Sammlung von Texten, die einen Ausschnitt einer Sprache repräsentieren, wobei vorausgesetzt wird, daß diese Sammlung zu einem bestimmten Zweck erstellt wird.

Untersucht werden soll die Frage, ob aus dem WWW linguistisch gut verwertbare Korpora gewonnen werden können, d. h. Korpora, mit deren Hilfe linguistische Fragestellungen beantwortet werden können. Die Motivation dafür liegt auf der Hand: Für viele Forschungsfragen gibt es bisher keine geeigneten Korpora, und die Erstellung von Korpora auf dem herkömmlichen Wege, d. h., ohne Zuhilfenahme von Quellen wie dem WWW, ist eine teure Angelegenheit.

Es besteht ein großer Bedarf an spezialisierten Korpora, um Untersuchungen in bestimmten Themenbereichen vornehmen zu können. Weiterhin werden große Korpora benötigt, denn das Problem der LNRE-Verteilung (Large Number of Rare Events) kann nur durch die Größe eines Korpus kompensiert werden. Dementsprechend erscheint das WWW als sinnvolle Lösung. Es enthält große Mengen Text zu den verschiedensten Themen in elektronischer Form, der leicht zugänglich ist und schnell und billig weiterverarbeitet werden kann.

Eine Methode, die sich zu diesem Zweck anbietet, ist von Baroni & Bernardini (2004)<sup>2</sup> entwickelt worden. Dabei wird die Internet-Suchmaschine Google<sup>3</sup> verwendet. Zuerst wird nach einigen wenigen für das Zielkorpus repräsentativen Suchtermen – *seeds* – gesucht, und die besten Treffer bilden heruntergeladen ein Anfangskorpus. Aus diesem werden durch den Vergleich der Wortfrequenzen mit denen eines Referenzkorpus neue charakteristische Suchterme extrahiert, die wieder als Eingabe für Google dienen etc. Auf diese Weise kann sehr schnell ein Spezialkorpus erstellt werden, dessen Umfang mit der Zahl der Iterationen des Algorithmus gut wählbar ist. Weitere Parameter, die das Zielkorpus beeinflussen, sind die Anzahl der Suchanfragen pro Iteration, die Zahl der Seeds, die kombiniert werden, um Suchanfragen zu bilden sowie die Zahl der Trefferseiten, die heruntergeladen werden sollen.

Als Beispiel für die Untersuchung der oben genannten Fragestellung soll ein Fach-/Spezialkorpus zum Thema „Sport“ erstellt werden. Dieses Sportkorpus kann in einer parallel laufenden Magisterarbeit zur Untersuchung der grammatischen Integration von Anglizismen im Deutschen dienen.

Ich werde im Vortrag das genannte Verfahren demonstrieren und erfolgreiche Anwendungen vorstellen. Ich komme zu dem Schluß, daß das Verfahren zur Erstellung des Sportkorpus ungeeignet ist und nur mit hochspezialisierten Domänen funktioniert.

---

<sup>1</sup>John Sinclair: *Corpus, Concordance, Collocation*. Oxford University Press. Oxford: 1991

<sup>2</sup>Baroni, Marco & Bernardini, Silvia: *BootCaT: Bootstrapping Corpora and Terms from the Web*.  
[http://sslmit.unibo.it/~baroni/lrec2004/bootcat\\_lrec.2004.pdf](http://sslmit.unibo.it/~baroni/lrec2004/bootcat_lrec.2004.pdf)

<sup>3</sup><http://www.google.com>