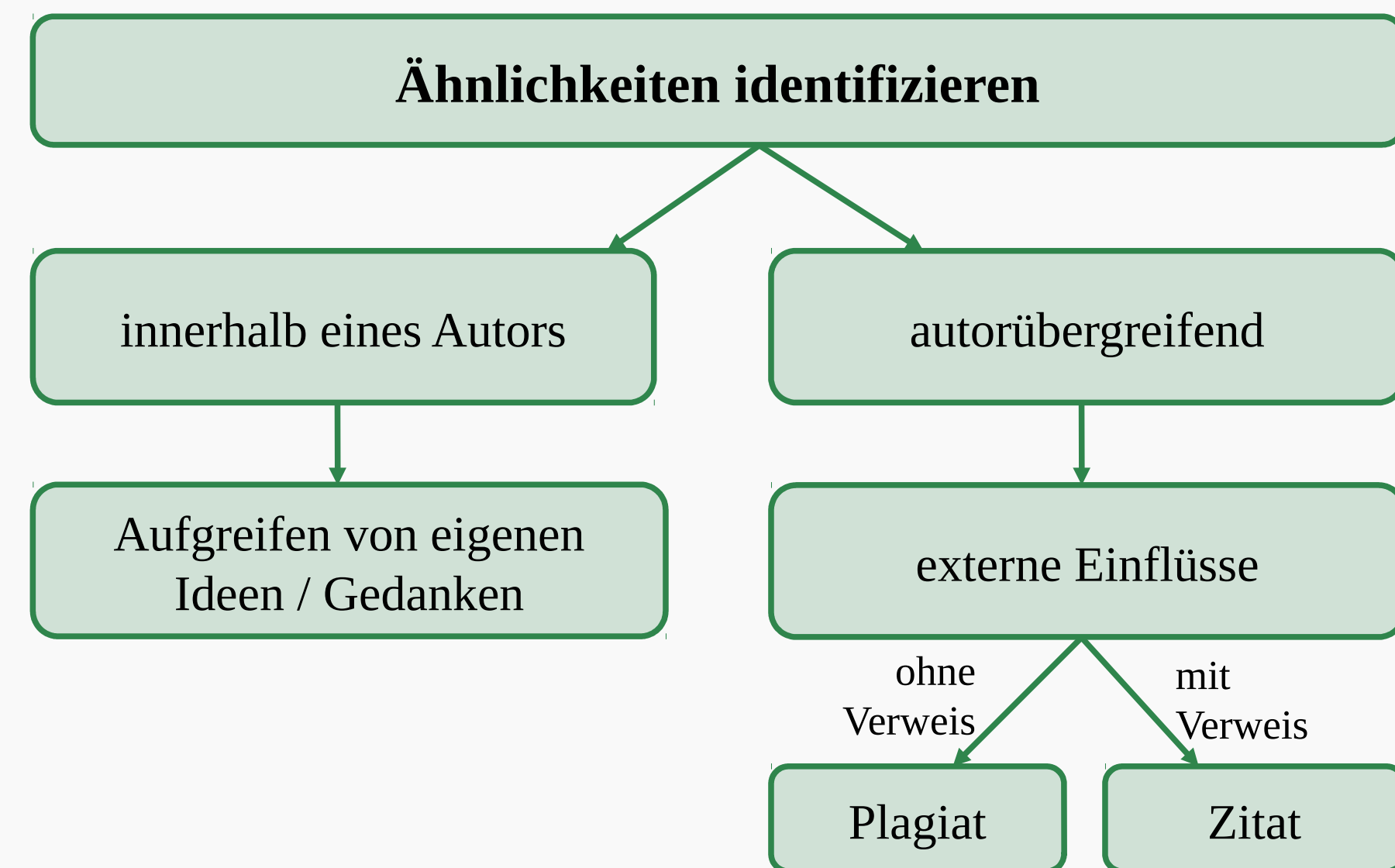


Sabine Ullrich¹, Daniel Bruder², Maximilian Hadersbeck¹¹Centrum für Informations- und Sprachverarbeitung, LMU München²Department of Computer Science and Technology, University of Cambridge

1. Ziel



- Entwicklung eines Tools zur Unterstützung des Philologen beim **Aufspüren von Einflüssen in textgenetischen Prozessen**
- Vorfilterung** von ähnlichen Textabschnitten mit Methoden des Machine Learning
- Gewähren von **Zeit** für intensivere Analysen

2. Anwendungsbereiche

- Digital Humanities** (Nachlass L. Wittgenstein: WittFind): Finden von Einflüssen in philosophischen Texten, interne als auch externe
- Plagiatsforschung**: Einflüsse aus externen Quellen, insbesondere nicht (direkt oder indirekt) zitierte Textstellen
- Politische oder religiöse Kontexte**: z.B. Ähnlichkeiten zwischen verschiedenen religiösen Texten verschiedener Religionen

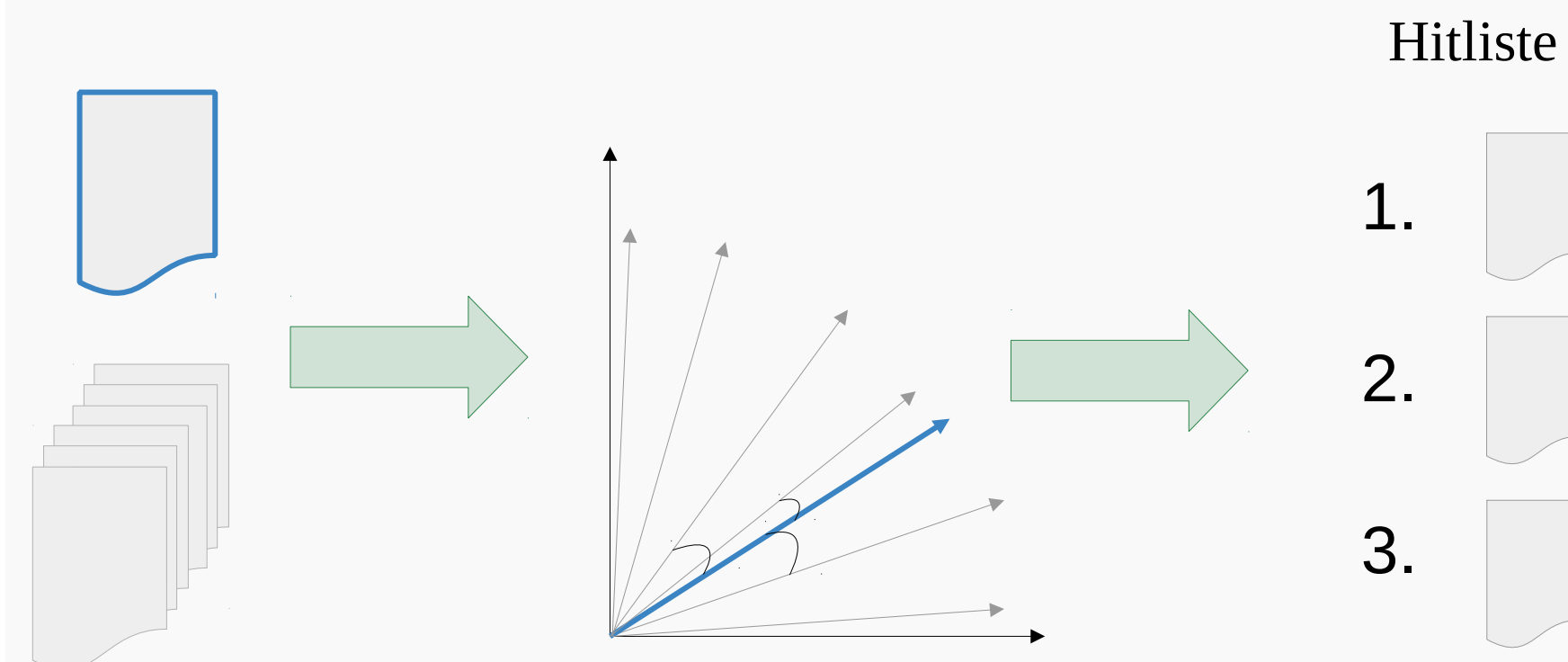
Kontakt

sabine.ullrich@campus.lmu.de
dmb77@cam.ac.uk
Maximilian.Hadersbeck@lmu.de

3. Methode

Berechnung der Ähnlichkeiten zwischen Textstellen:

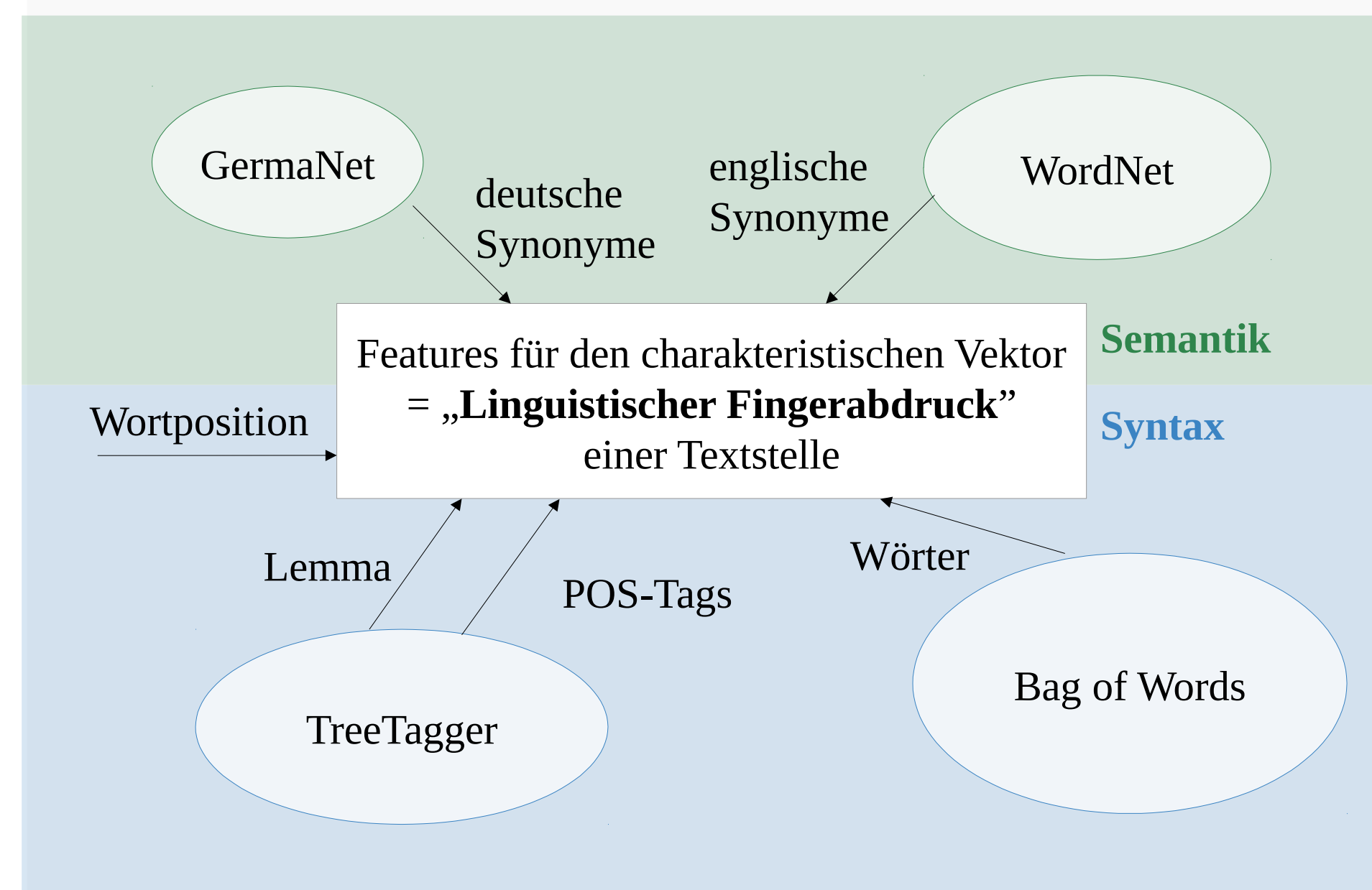
- Bestimmung der **charakteristischen Vektoren** einer (zu prüfenden) Sammlung von Textstellen (offline vorberechnet für effiziente und schnelle Suche)
- Bestimmung des charakteristischen Vektors der Eingabe
- Filtern der Dokumente mit dem kleinsten **Cosinus Abstand**
- Ausgabe in einer Hitliste



Das **Ähnlichkeitsmaß** berechnet sich wie folgt:
 $\text{Ähnlichkeit} = 1 - \text{Cosinus Abstand}$

0% = **keine Übereinstimmung** zwischen den Textstellen
100% = die Textstellen sind **identisch**

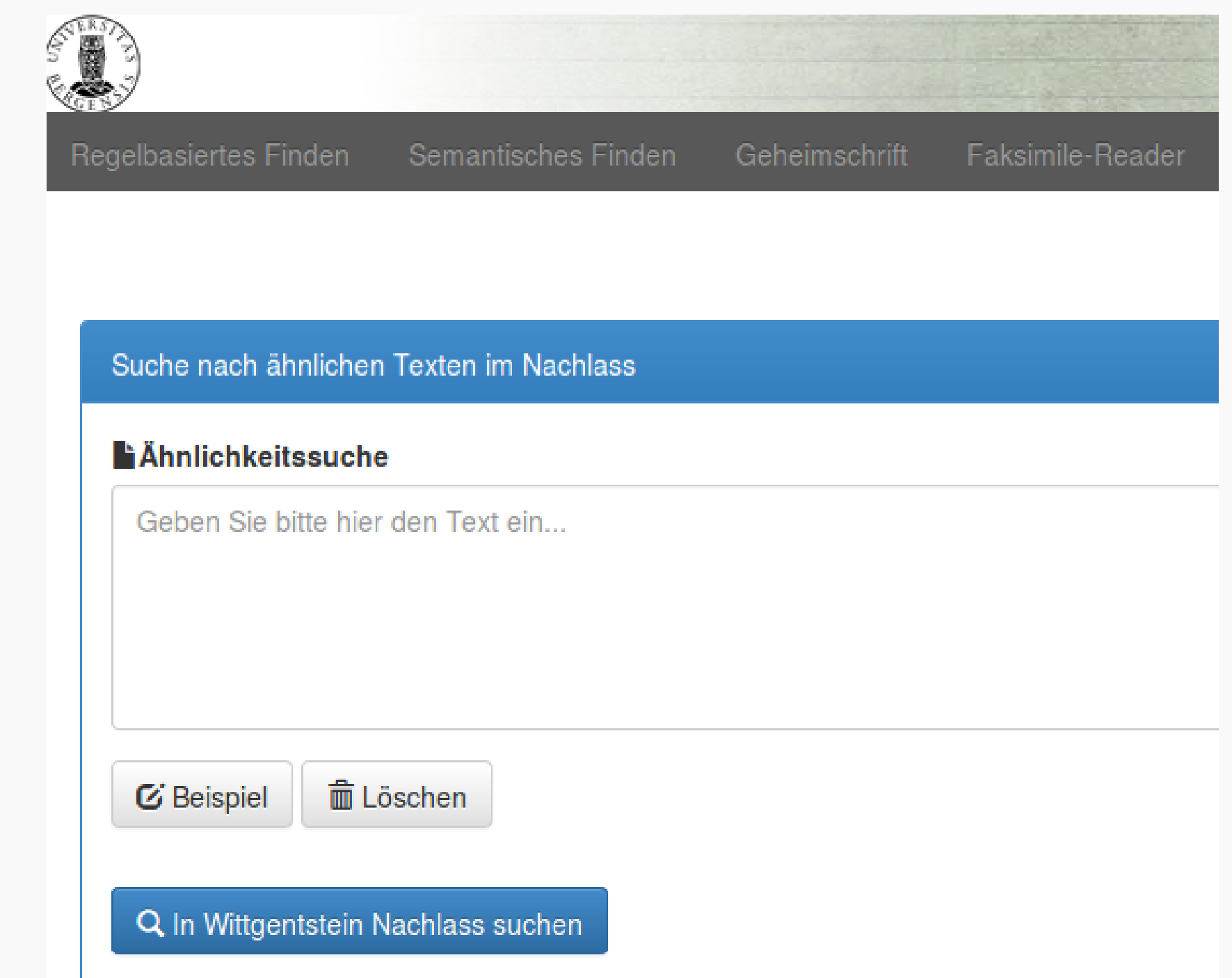
4. Vektorbestimmung



5. Integration in WittFind

Sobald das Tool zur Ähnlichkeitsbestimmung finalisiert ist, wird es in der Suchmaschine WittFind die **Wittgenstein Advanced Search Tools** (WAST) erweitern:

<http://wittfind.cis.lmu.de/>



Referenzen

Hadersbeck, Maximilian / Pichler, Alois / Fink, Florian / Gjesdal, Øyvind, L. (2014): *Wittgenstein's Nachlass: WITTFind and Wittgenstein Advanced Search Tools* (WAST). DATeH. Madrid.

Hamp, Birgit / Feldweg, Helmut, et al. (1997): „Germanet-a lexical-semantic net for german“. In: *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 9–15.

Henrich, Verena / Hinrichs, Erhard (2010): „Gernedit-the germanet editing tool“. In: *ACL (System Demonstrations)*, 19–24. Citeseer.

Pichler, Alois / Krüger, Heinz W. / Smith, D. / Bruvik, Tone / Lindebjerg, Anne / Olstad, Vemund (Hrsg.) (2009): *Wittgenstein Source Bergen Facsimile* (BTE). Wittgenstein Source Bergen.

Schmid, Helmut (1995): „Improvements in part-of-speech tagging with an application to german“. In: *Proceedings of the ACL sigdat-workshop*. Citeseer.

Schmid, Helmut (1995): *Probabilistic part-of-speech tagging using decision trees*. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 43:28–37, 1995.

Wittgenstein, Ludwig / Nedo, Michael (Hrsg.) (1996): *Wiener Ausgabe*. Band 1-5.