

# **Ein quantitativer Vergleich zwischen UD und SUD**

**anhand der Textkomplexitätsmaße**

Projektbericht zur Lehrveranstaltung

**DH 1: Sprache und Text (SS 2025)**

im Studiengang

**Digitale Geistes- und Sozialwissenschaften (BA)**

an der Friedrich-Alexander-Universität Erlangen-Nürnberg

vorgelegt von

**Xinyao Lu**

Computerlinguistik (8. FS) / Germanistik (8. FS)

Matrikelnummer: 23008498

Kontakt: xinyao.lu@fau.de

und

**Julian Vignisson**

Data Science (8. FS)

Matrikelnummer: 23003258

Kontakt: julian.vignisson@fau.de

Betreuer(in):

Andreas Blombach, Marianna Gracheva

Erlangen, 12.9.2025

# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>2</b>
1.1	Das SUD-Annotationsschema . . . . .	2
1.2	Maße der syntaktischen Komplexität . . . . .	4
<b>2</b>	<b>Zielsetzung</b>	<b>4</b>
2.1	Die Hypothesen . . . . .	4
<b>3</b>	<b>Methodik</b>	<b>5</b>
3.1	Auswahl des Datensatzes . . . . .	5
3.2	Annotation und Berechnung der Komplexitätsmaße . . . . .	6
<b>4</b>	<b>Ergebnisse</b>	<b>8</b>
4.1	Deskriptive Ergebnisse . . . . .	8
4.2	Hypothesentests . . . . .	10

# 1 Einleitung

## 1.1 Das SUD-Annotationsschema

*Surface Syntactic Universal Dependencies* (**SUD**) ist ein alternatives Annotat ionsschema zu *Universal Dependencies* (**UD**), ein sprachübergreifendes System zur Annotation der grammatischen Strukturen. SUD priorisiert die syntaktischen Funktionen der Satzglieder, indem es die **Funktionswörter** (Hilfsverben, Präpositionen etc.) anstelle von **Inhaltswörtern** (Substantiven, Adjektiven und lexikalischen Verben etc.) als Köpfe der Konstruktion zuweist. Nach dieser Umstrukturierung werden das Subjekt und alle adverbialen Ergänzungen bei SUD dem Hilfsverb zugeordnet, während die Objekte beim lexikalischen Verb verbleiben. Darüber hinaus hängt SUD unter dem Prinzip der **Minimierung der Dependenzlänge** (*dependency length minimization principle*) die nebengeordneten Dependents (inkl. Konjunkten) an deren jeweiligen am nächsten stehenden Nachbar; nur die erste Dependent hängt direkt vom Kopf ab. Dadurch entsteht bei SUD eine Kettenstruktur vom Kopf bis zur letzten Dependent. Beide dieser Änderungen – die Verteilung der Dependents auf Hilfs- und lexikalische Verben sowie das Anhängen der Dependents an den nächsten Nachbarn – sollten prinzipiell zu einer **tieferen** und **dezentralisierten** syntaktischen Struktur führen als die verhältnismäßig flachere und zentralisierte von UD. [Gerdes et al. 2018]

Die beiden folgenden Beispiele veranschaulichen die Änderungen von SUD. Im ersten Beispiel (s. Abb. 1 und Abb. 2: Satzwurzel in UD und SUD) wird die Wurzel des Satzes vom lexikalischen Verb *like* zum Hilfsverb *did* verlagert. Im zweiten Beispiel (s. Abb. 3 und Abb. 4: Konjunkten in UD und SUD) wird der dritte Konjunkt *radio* von SUD an den benachbarten *podcasts* angehängt. (vgl. bei UD an den ersten Konjunkt *shows*)

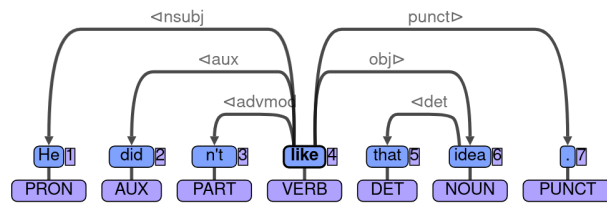


Abbildung 1: UD: Lexikalisches Verb *like* als Wurzel

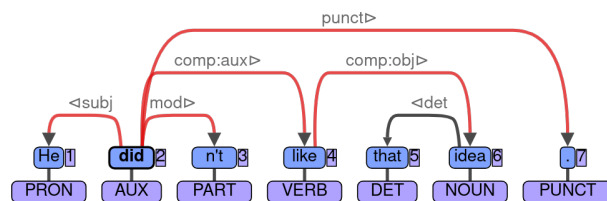


Abbildung 2: SUD: Hilfsverb *did* als Wurzel

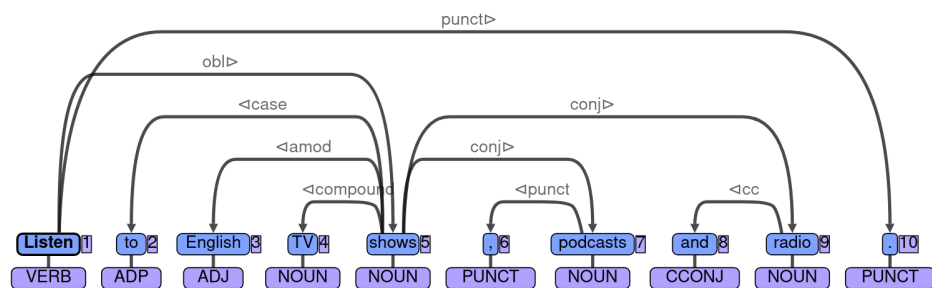


Abbildung 3: Konjunkten unter UD

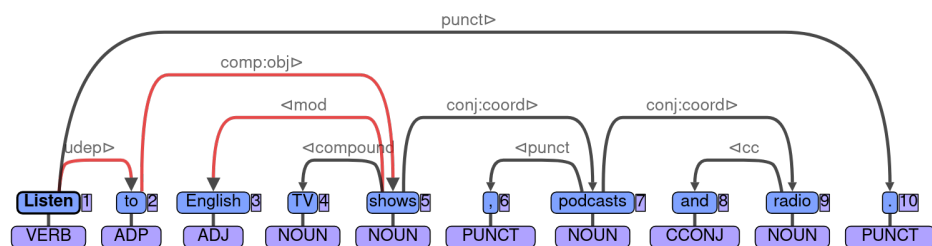


Abbildung 4: Konjunkten unter SUD

## 1.2 Maße der syntaktischen Komplexität

Die Abhängigkeitsstruktur eines Satzes (manchmal auch **syntaktische Bäume** genannt) lässt sich mit **Komplexitätsmaßen** von der Graphentheorie analysieren, wenn man die Tokens als Knoten und die Abhängigkeitsrelationen als gerichtete Kanten in einem Graph behandelt, z. B. die **Durchschnittliche Abhängigkeitsdistanz** (*Average Dependency Distance*, [Oya 2021]) berechnet die Entfernung von jeder Dependent zu ihrem Kopf. Nähere Auslegungen über die Komplexitätsmaße und ihre Anwendung bei der Abhängigkeitsanalyse sind in der Sektion **3.2 Annotation und Berechnung der Komplexitätsmaße** zu finden.

## 2 Zielsetzung

### 2.1 Die Hypothesen

Wie erörtert in der Einleitung, wird erwartet, dass das SUD-Schema im Vergleich zum UD-Schema eine **tiefere** und **dezentralisierte** syntaktische Struktur erzeugen sollte. Diese erwartete Differenz lässt sich in drei konkrete Hypothesen aufteilen. In diesem Seminarprojekt wurden die Hypothesen empirisch überprüft.

**Hypothese 1 (H<sub>1</sub>):** *Die syntaktischen Bäume unter dem SUD-Schema seien **tiefer** als die unter dem UD-Schema.*

**Hypothese 2 (H<sub>2</sub>):** *Das SUD-Schema solle die durchschnittliche **Abhängigkeitsdistanz reduzieren** und somit das Prinzip der Minimierung der Abhängigkeitslänge besser erfüllen.*

**Hypothese 3 (H<sub>3</sub>):** *Das SUD-Schema solle die **Überzentralität** der Inhaltswörter reduzieren, indem es die Dependents zwischen Inhaltswörtern und Funktionswörtern aufteilt.*

## 3 Methodik

### 3.1 Auswahl des Datensatzes

Eine englischsprachige Textsammlung, die aus ungefähr 1.500 kurzen Texten von 6 Schwierigkeitsniveau (A1 - C2) nach Gemeinsamen Europäischem Referenzrahmen für Sprachen (**GER**, engl. *Common European Framework of Reference*) besteht, wurde als Datensatz zur Durchführung des Experiments ausgewählt.<sup>1</sup> Nach Metadaten dieses Datensatzes seien die Texte aus einigen zuverlässigen Online-Lernplattformen für die englische Sprache ausgewählt worden. Das Schwierigkeitsniveau wurde von den Lernplattformen übernommen oder durch ein Online-Tool *Text Inspector*<sup>2</sup> automatisch bewertet. Er fasst Texte aus umfangreichen Themengebieten, wie alltäglichem Dialog, Zeitungsbeiträgen und Geschäftskommunikation, um. Diese sechsstufige Textsammlung solle im Experiment zur Überprüfung beitragen, ob der Unterschied zwischen UD und SUD bei verschiedenen Schwierigkeitsniveaus **konsistent** sei.

Die **Textmenge** von 6 Niveaus ist im Datensatz insgesamt gleichmäßig verteilt: Mit einem Mittelwert von 250 variiert die Menge von 200 bis zu 280. Jedoch weisen Texte aus verschiedenen Niveaus auf erheblichen Unterschied bei der **Textlänge** hin: Während der durchschnittliche Wert von den steigenden Niveaus unter 300 Tokens (A1 119,8 und A2 270,5 Tokens pro Text) liegt, erreichen manche fortgeschrittene Texte eine Länge von über 1.000 Tokens. (Mittelwert von C1 786,2 und C2 804,4 Tokens pro Text)

Die **Durchschnittliche Tokenzahl pro Satz** wird häufig als ein einfaches Maß zur Textkomplexität verwendet. In diesem Projekt haben wir dieses Maß berechnet, um die **Kohärenz** in jedem Schwierigkeitsniveau abzuschätzen. Das Ergebnis (s. Abb. 5: Durchschnittliche Tokenzahl pro Satz nach Niveau) zeigt deutliche Abgrenzung zwischen allen 6 Niveaus und die durchschnittliche Tokenzahl nimmt kontinuierlich mit steigendem Niveau zu. Deswegen gehen wir davon aus, dass die Zuordnung der Schwierigkeitsniveaus vom Datensatz **generell zuverlässig** sein sollte.

---

<sup>1</sup><https://www.kaggle.com/datasets/amontgomerie/cefr-levelled-english-texts/data>

<sup>2</sup><https://textinspector.com/help/tu-lexical-profile/>

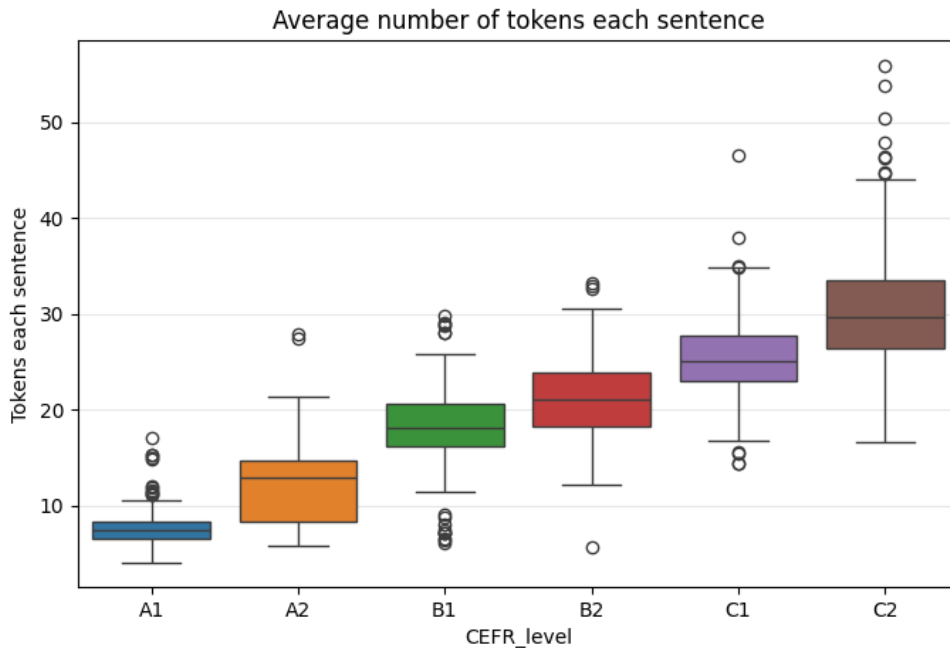


Abbildung 5: Durchschnittliche Tokenzahl pro Satz nach Niveau

Bei der Datenvorverarbeitung wurden nur die Kodierungsfehler (z. B. das Unicode-BOM-Zeichen *uff*, falsche Dekodierung der Umlaute) entfernt bzw. korrigiert. Die non-ASCII Zeichen wie Diakritika, Mathematik- und Währungszeichen wurden in den Texten überlassen, da die neuesten Taggingsmodellen solche Sonderzeichen ohne Schwierigkeit taggen können. Sechs Texte aus dem A1 Niveau wurden ausgeschlossen, weil sie aus keinen grammatisch vollständigen Sätzen bestehen und sich nicht für den Syntactagging eignen. Es handelt sich bei denen um Speisekarten, Wörterbucheinträge, Werbungen und Visitenkarten. Alle anderen Texte aus dem Datensatz wurden im Original in das Experiment übernommen.

### 3.2 Annotation und Berechnung der Komplexitätsmaße

In diesem Projekt wurde das NLP-Toolkit *Stanza* (Qi et al. 2020) genutzt, um den gegebenen Datensatz syntaktisch im UD-Format zu annotieren. Stanza stellt vor-trainierte Modelle bereit, die jedem Wort im Satz eine Wortart (UPOS/XPOS) und morphologische Merkmale zuweisen und **Abhängigkeitsrelationen** zwischen

den Wörtern identifizieren. Für die Analyse wurde eine Stanza-Pipeline in Englisch konfiguriert und auf jeden Satz angewendet. Als Ergebnis erhielt jeder Satz einen gerichteten Abhängigkeitsbaum mit einem Wurzelknoten (in der Regel das finite Prädikat) und Kanten zu seinen Dependents, jeweils etikettiert mit einem Dependenzlabel. Die mit Stanza erzeugten Annotationen wurden anschließend in das *CoNLL-U*-Format<sup>3</sup> exportiert. **CoNLL-U** ist ein standardisiertes tabellarisches Textformat zur Speicherung von UD-annotierten Sätzen. Durch den Export in das CoNLL-U-Format lagen die syntaktischen Analysen in einer einheitlichen und weiterverarbeitbaren Form vor. Nachdem der Datensatz von Stanza getaggt wurde, wurden die UD-Bäume in das SUD-Schema konvertiert. Die Konvertierung erfolgte mit dem Tool *Grew*<sup>4</sup>, das die SUD-Bäume ebenfalls im CoNLL-U-Format speichert. Dies dient als Grundlage für nachfolgende Verarbeitungsschritte. Für die Auswertung kam das Python-Paket `textcomplexity` [Proisl 2017] zum Einsatz, das eine Reihe von Maßzahlen zur linguistischen Komplexität implementiert. Insbesondere wurden diese 4 Dependenz-basierten Maße auf Satzebene pro Text ermittelt:

- **Baumtiefe** (engl. *Longest Shortest Path*): die maximale Entfernung (Anzahl Kanten) vom Wurzelknoten zum am weitesten entfernten Token im syntaktischen Baum. Diese Größe spiegelt die hierarchische Tiefe des Satzes wider – ein höherer Wert deutet auf komplexere, tief geschachtelte Strukturen hin.
- **Durchschnittliche Abhängigkeitsdistanz** (*Average Dependency Distance*): ein Maß nach [Oya 2021], das den mittlere linearen Abstand zwischen Köpfen und Dependents über alle Abhängigkeiten eines Satzes beschreibt. Diese Metrik quantifiziert, wie weit auseinander im Satz zusammengehörige Satzteile typischerweise stehen. Größere durchschnittliche Distanzen können auf komplexere Satzstrukturen hindeuten, da die zusammenhängenden Satzglieder weiter über den Satz verteilt sind.
- **Closeness-Zentralisierung**: ein Maß nach [Freeman 1978] für die Zentralisierung des Abhängigkeitsgraphen basierend auf der *Closeness Centrality* jedes Knotens. Konkret wird betrachtet, wie nahe (in graphentheoretischer Distanz) alle Tokens im Mittel dem Zentrum (der Wurzel) sind, verglichen mit

---

<sup>3</sup>Spezifikation: <https://universaldependencies.org/format.html>.

<sup>4</sup><http://grew.fr>



einer vollständig sternförmigen Struktur. Unter einer sternförmigen Struktur versteht man einen Graphen mit genau einem zentralen Knoten, der direkt mit allen übrigen Knoten verbunden ist, während zwischen den übrigen Knoten keine Kanten bestehen. Ein Wert von 1 würde bedeuten, dass ein Knoten (hier typischerweise die Wurzel) direkt mit allen anderen verbunden ist; niedrigere Werte zeigen an, dass die Abhängigkeitsstruktur „verteilter“ ist und die Wörter weiter vom Satzkern entfernt liegen.

- **Outdegree-Zentralisierung:** ebenfalls ein Zentralisierungsmaß nach Freeman, das hier auf der Verteilung der ausgehenden Kanten (*outdegree*) pro Knoten basiert. Dieses Maß liegt ebenfalls zwischen 0 und 1 und erreicht 1, wenn ein einzelner Knoten alle Dependents besitzt (d. h. alle anderen Wörter direkt unter sich hat, was auf eine maximale Zentralisierung um die Wurzel hindeutet). Niedrigere Werte bedeuten, dass die Abhängigkeiten auf mehrere Knoten verteilt sind.

Die Berechnung dieser Kennzahlen erfolgte automatisiert mit `textcomplexity`, das die Dependenzstrukturen (eingelesen im CoNLL-U-Format) zu Graphen umwandelt (unter Nutzung von *NetworkX*) und dann die vordefinierten Formeln anwendet. Für jeden Text wurden die Kennwerte zunächst satzweise berechnet und anschließend auf Textebene gemittelt, sodass pro Text ein durchschnittlicher Wert und Metrik vorlag. Dadurch lagen pro Text jeweils ein UD- und ein SUD-Wert für jede Metrik vor, was einen direkten Vergleich ermöglichte.

## 4 Ergebnisse

### 4.1 Deskriptive Ergebnisse

Insgesamt zeigen sich die gemessenen dependenzbasierten Metriken über alle geprüften Sprachniveaus hinweg **systematische Unterschiede** zwischen UD und SUD. (s. Abbildung 6: Metriken von UD und SUD) Im Verlauf von Niveau A1 bis C2 nehmen erwartungsgemäß die Tiefe der Bäume und die durchschnittliche Abhängigkeitsdistanz zu. Gleichzeitig nehmen die Zentralisierungsmaße mit steigendem Niveau ab. Die höhere Syntaxkomplexität von fortgeschrittenen Texten äußert

sich in **tieferen Bäumen**, **längeren Abhängigkeitsdistanzen** und **niedrigeren Zentralisierungswerten**. Diese allgemeinen Trends zeigen sich sowohl in UD als auch in SUD. Allerdings unterscheiden sich die absoluten Werte zwischen den beiden Annotationsschemata.

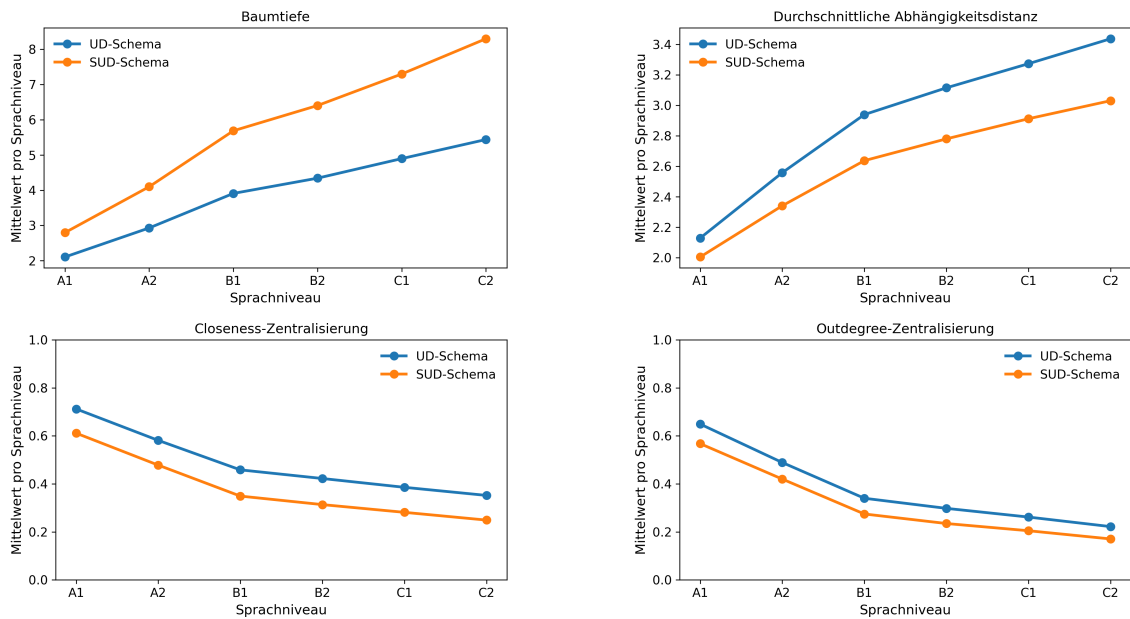


Abbildung 6: Vergleich der syntaktischen Metriken zwischen UD- und SUD-Schema über die GER-Niveaus (A1–C2): die Baumtiefe (oben links), die durchschnittliche Abhängigkeitsdistanz (oben rechts), die Closeness-Zentralisierung (unten links) und die Outdegree-Zentralisierung (unten rechts)

- **Baumtiefe:** In SUD sind die Bäume über alle Niveaustufen hinweg konsistent tiefer als in UD, dieser Abstand vergrößert sich bei höheren Niveaus.
- **Durchschnittliche Abhängigkeitsdistanz:** Dagegen ist die durchschnittliche Abhängigkeitsdistanz in SUD durchweg geringer als in UD. Das heißt, syntaktisch verbundene Wörter stehen im SUD-Baum tendenziell näher beieinander als im UD-Baum; Die Behauptung von SUD, dass SUD das Prinzip von Minimierung der Dependenzlänge besser erfüllt, erhält an dieser Stelle einen empirischen Beweis. Auch dieser Unterschied zeigt sich konsistent auf allen Schwierigkeitsniveaus.

- **Zentralisierungsmaße:** Auch die beiden Zentralisierungsmaße weisen klare Differenzen zwischen UD und SUD auf. Closeness-Zentralisierung und Outdegree-Zentralisierung fallen in SUD über alle Niveaus hinweg systematisch niedriger aus als in UD. Dies bedeutet, dass die syntaktischen Abhängigkeiten im SUD-Schema insgesamt gleichmäßiger auf mehrere Knoten verteilt sind. Die beiden hauptsächlichen Änderungen von SUD - Verteilung der Dependents auf Inhalts- und Hilfsverben und neue Zuordnung der nebengeordneten Dependents (vgl. Sektion 1.1 Das SUD-Annotationsschema) - sollten eine wichtige Rolle bei der Senkung der Zentralisierungswerte gespielt haben.

## 4.2 Hypothesentests

Über die beobachtete Differenz zwischen UD und SUD wurden die statistische Signifikanz folgendermaßen überprüft. Da sich ein  $UD_i$ - und ein  $SUD_i$ -Wert jeweils auf den gleichen Text  $i$  beziehen, sind die Daten paarweise abhängig. Deshalb nutzen wir einen gepaarten t-Test, welcher als Signifikanztest die Mittelwerte zweier gepaarter Messungen vergleicht. Dabei wird die Nullhypothese  $H_0$  geprüft, dass die  $\tilde{d}$  Differenz  $d_i = SUD_i - UD_i$  unerheblich ist, d. h. es gibt keinen systematischen Unterschied zwischen UD und SUD. Voraussetzung hierfür ist, dass die Differenzen einer Normalverteilung folgen. Um dies zu überprüfen, wurde die Verteilung der Differenzwerte ( $SUD - UD$ ) jeder Metrik untersucht. Hierzu wurden QQ-Plots erstellt (Abb. 7: QQ-Plots), die keine deutlichen Abweichungen von der Geraden zeigten.<sup>5</sup> Die Differenzen lassen sich also durch eine Normalverteilung annähern.

---

<sup>5</sup>`scipy.stats` s. <https://docs.scipy.org/doc/scipy/reference/stats.html>

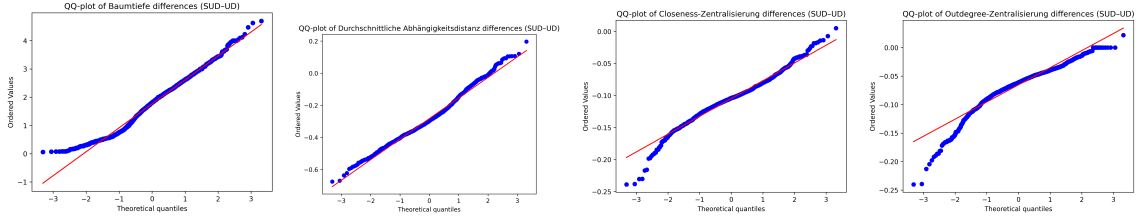


Abbildung 7: QQ-Plots zur Überprüfung der Normalverteilung der Differenzen (SUD – UD) für die vier untersuchten Metriken. Die Plots wurden mit der Funktion `probplot` aus `scipy.stats` erzeugt. Die roten Linien stellen die theoretische Normalverteilung dar, die blauen Punkte die beobachteten Quantile.

Für die **1. Hypothese**, dass die syntaktischen Bäume unter SUD tiefer sind als unter UD, erhalten wir ein klares Ergebnis im  $t$ -Test. Wir erhalten  $t = 16.863$  und  $p = 3.084 \times 10^{-24}$ . Die SUD-Bäume sind im Mittel deutlich tiefer als die UD-Bäume. Der hohe  $t$ -Wert und der extrem kleine  $p$ -Wert zeigen, dass dieser Unterschied hochsignifikant ist. Damit können wir die 1. Hypothese bestätigen.

Für die **2. Hypothese**, die durchschnittliche Dependenzlänge, erhalten wir  $t = -17.053$  und  $p = 1.783 \times 10^{-24}$ . Die durchschnittliche Dependenzlänge ist unter SUD signifikant kleiner als unter UD. Das negative Vorzeichen kommt darauf zurück, dass SUD-Werte niedriger als UD-Werte sind. Damit können wir die 2. Hypothese bestätigen.

Für die **3. Hypothese**, die Zentralität der Inhaltswörter, haben wir zwei Maße berechnet und erhalten zwei Testergebnisse. Für die Closeness-Zentralisierung gilt  $t = -43.237$  und  $p = 2.226 \times 10^{-46}$ . Für die Outdegree-Zentralisierung erhalten wir  $t = -21.635$  und  $p = 9.642 \times 10^{-30}$ . Obwohl sie unterschiedliche Aspekte der Zentralität erfassen, führen sie zum gleichen Ergebnis: Unter SUD sind die Zentralisierungswerte auf Satzebene signifikant niedriger als unter UD. Eine wohl mögliche Ursache zu den gesunkenen Zentralisierungswerten könnte sein, dass unter SUD-Schema die zentralen Knoten (beispielsweise die Prädikaten) weniger stark fungieren, während mehr Knoten (Präpositionen, Hilfsverben) auch eine Kopfrolle übernehmen. Damit betrachten wir das Textergebnis von den beiden Zentralisierungsmetriken als einen indirekten Beweis zur 3. Hypothese.

## Ausblick

Die vorliegenden Ergebnisse sind insgesamt als sehr aussagekräftig einzuschätzen. Die Analysen beruhen auf einer großen Datenmenge, die alle sechs GER-Niveaus (A1–C2) abdeckt, und zeigen konsistente Effekte über alle Niveaus hinweg. Damit ist die zentrale Aussage, dass das SUD-Schema im Vergleich zu UD tiefere Bäume erzeugt und die durchschnittliche Dependenzlänge reduziert, abgesichert.

Auch für die Verringerung der Überzentralität von Inhaltswörtern haben wir einen Beweis gefunden. Allerdings liefern die hier verwendeten Zentralisierungsmaße jeweils einen aggregierten Wert auf die Satzebene und erlauben keinen direkten Rückschluss auf die Zentralität einzelner Wörter. Diese Beziehung wurde im Rahmen des Projekts nicht weiter untersucht. Eine mögliche Erweiterung bestünde darin, die `networkx`<sup>6</sup>-Repräsentation als `DiGraph` zu nutzen, um die Zentralität auf Knoten- bzw. Tokenebene zu berechnen und so die 3. Hypothese präziser zu überprüfen.

Darüber hinaus könnte in einer weiterführenden Untersuchung geprüft werden, ob die Stärke der beobachteten Effekte zwischen den einzelnen GER-Niveaus variiert. Methodisch ließen sich hierfür etwa lineare gemischte Modelle einsetzen, die Interaktionen zwischen Schema und Sprachniveau explizit modellieren.

## Bemerkungen

- **Zur Aufgabeteilung:** Xinyao Lu war in der Zusammenarbeit für die **linguistische** Theorie verantwortlich, während Julian Vignisson seine Fachkenntnisse in **Statistik** einbrachte.
- Xinyao Lu beschäftigte sich mit der Feststellung der **Hypothesen**, der **Auswahl** und **Vorverarbeitung** des Datensatzes. Er hat die Sektionen 1, 2 und 3.1 vom Projektbericht geschrieben.
- Julian Vignisson kümmerte sich um den **Tagging** und SUD-Konvertierung des Datensatzes, die Berechnung der **Metriken** und die **Signifikanztests**. Er hat 3.2 und 4 im Projektbericht geschrieben.

---

<sup>6</sup><https://networkx.org/>

- Die gesamten Skripte des Seminarprojekts sind in diesem GitHub-Repository zu finden: <https://github.com/cometbridge1998/SUDcom.git>

## Literatur

- Freeman, Linton C. (1978): Centrality in social networks conceptual clarification. In: *Social Networks* 1 (3), S. 215–239. URL: <https://www.sciencedirect.com/science/article/pii/0378873378900217>.
- Gerdes, Kim, Bruno Guillaume, Sylvain Kahane und Guy Perrier (Nov. 2018): SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In: *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*. Hrsg. von Marie-Catherine de Marneffe, Teresa Lynn und Sebastian Schuster. Brussels, Belgium: Association for Computational Linguistics, S. 66–74. URL: <https://aclanthology.org/W18-6008/>.
- Oya, Masanori (Nov. 2021): Three Types of Average Dependency Distances of Sentences in a Multilingual Parallel Corpus. In: *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*. Hrsg. von Kaibao Hu, Jong-Bok Kim, Chengqing Zong und Emmanuele Chersoni. Shanghai, China: Association for Computational Linguistics, S. 652–661. URL: <https://aclanthology.org/2021.paclic-1.69/>.
- Proisl, Thomas (2017): *textcomplexity*. <https://github.com/tsproisl/textcomplexity>.
- Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton und Christopher D. Manning (2020): Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. URL: <https://nlp.stanford.edu/pubs/qi2020stanza.pdf>.