



LATIC

LATIC – Ein linguistisches Analysetool für Itemcharakteristika

Dokumentation (Version 1.0.0)

Nadine Cruz Neri¹ & Florian Klückmann²

¹ Fakultät der Erziehungswissenschaft, Arbeitsbereich Pädagogische Psychologie, Universität
Hamburg

² Freiberuflicher Softwareentwickler

Hamburg, 14. Mai 2021

Autor:innennotiz

Ein besonderer Dank gilt Jamie Böhmer für die Erstellung des Logos für LATIC sowie die Bereitstellung weiterer Designelemente.

Bei Fragen oder Anmerkungen zur Dokumentation sowie zur Software LATIC wenden Sie sich per Mail an uns (hello@latic.software) oder hinterlassen Sie einen Kommentar in GitHub unter <https://github.com/florianklueckmann/LATIC/issues>.

Empfohlene Zitation: Cruz Neri, N., & Klückmann, F. (2021). *LATIC – Ein linguistisches Analysetool für Itemcharakteristika* (Version 1.0.0) [Computer Software]. <https://github.com/florianklueckmann/LATIC>

Inhaltsverzeichnis

1 Einleitung	5
2 Technische Details	5
3 Sprachen	6
4 Anleitung	6
4.1 Herunterladen und Starten der Anwendung	6
4.2 Anwendung	7
4.3 Allgemeine Hinweise	8
5 Funktionen von LATIC	9
5.1 Analyse auf Wortebene	9
5.1.1 Adjektive	9
5.1.2 Adpositionen	9
5.1.3 Adverbien	11
5.1.4 Determinanten	11
5.1.5 Eigennamen	11
5.1.6 Hilfsverben	11
5.1.7 Interjektionen	11
5.1.8 Konjunktionen	12
5.1.9 Nomen	12
5.1.10 Partikel	12
5.1.11 Pronomen	13
5.1.12 Satzzeichen	13

5.1.13 Symbole	13
5.1.14 Unbekannt oder Ungewiss	13
5.1.15 Verben.....	13
5.1.16 Wortlänge.....	14
5.1.17 Zahlwörter.....	14
5.2 Analyse auf Satzebene.....	15
5.2.1 Satzanzahl	15
5.2.2 Satzlänge	15
5.3 Analyse auf Textebene	16
5.3.1 Annotiertes Item.....	16
5.3.2 Lesbarkeitsindices.....	16
5.3.3 Lexikvarianz	18
5.3.4 Wortanzahl.....	19
6 Evaluation.....	19
7 Literatur.....	22

LATIC – Ein linguistisches Analysetool für Itemcharakteristika

Die Rolle linguistischer Itemcharakteristika für das Leseverstehen und die Leistung von Lernenden gewinnt zunehmend an Aufmerksamkeit in der Forschung. Dies geschieht in verschiedenen Bereichen wie der Mathematik (Shaftel et al., 2006), den Naturwissenschaften (Cruz Neri et al., 2021) und dem Lesen im Allgemeinen (Heppt et al., 2015). Das Ziel ist es zu untersuchen, welche linguistischen Itemcharakteristika das Verstehen eines Textes oder eines Items erleichtern oder hemmen können (White, 2012). Zu diesem Zweck ist es unerlässlich, Items und Texte hinsichtlich verschiedener linguistischer Itemcharakteristika zu analysieren.

LATIC ermöglicht die Analyse und Auszählung verschiedener linguistischer Itemcharakteristika sowohl im Englischen als auch im Deutschen. Das Akronym LATIC steht dabei für „Linguistisches Analysetool für Itemcharakteristika“. Zur Annotation linguistischer Itemcharakteristika wird das Natural Language Processing Tool namens Stanford CoreNLP 4.2.1 (Manning et al., 2014) verwendet.

2 Technische Details

LATIC ist eine Java-Anwendung, die unter der Verwendung des Stanford CoreNLP 4.2.1 (Manning et al., 2014) die Analyse und das Auszählen von linguistischen Itemcharakteristika ermöglicht. Das Stanford CoreNLP 4.2.1 (Manning et al., 2014) nutzt das sogenannte Natural Language Processing (siehe Chowdhury, 2005), um das Annotieren von Wörtern und Symbolen zu ermöglichen. Es orientiert sich dabei am Universal Dependencies Project (2014–2020). LATIC berechnet und zählt zusätzlich weitere Itemcharakteristika aus. Darunter fallen unter anderem die Wortanzahl, die Anzahl von Sätzen sowie die Wortlänge.

Die Geschwindigkeit der Analyse von Items in LATIC ist abhängig von (1) der Länge des eingefügten Items, (2) der ausgewählten Itemcharakteristika sowie (3) der

Leistungsfähigkeit des Computers. Die Analyse nimmt in der Regel nur wenige Sekunden in Anspruch.

LATIC kann unter Windows, Linux und macOS verwendet werden. Zur Nutzung von LATIC empfehlen wir einen Arbeitsspeicher von mindestens 8 GB.

3 Sprachen

LATIC ermöglicht unter der Verwendung des Stanford CoreNLP 4.2.1 (Manning et al., 2014) die Analyse und das Auszählen linguistischer Itemcharakteristika im Englischen und im Deutschen. Das Stanford CoreNLP 4.2.1 (ebd.) ermöglicht das Annotieren weiterhin in den Sprachen (1) Arabisch, (2) Chinesisch, (3) Französisch und (4) Spanisch. Eine Implementierung dieser Sprachen in LATIC ist möglich. Jedoch ist dafür die Unterstützung von Personen nötig, die eine dieser Sprachen zumindest rezeptiv auf sehr gutem Niveau beherrschen. Bei Interesse einer Zusammenarbeit, treten Sie gerne mit der Erstautorin in Kontakt.

4 Anleitung

4.1 Herunterladen und Starten der Anwendung

Um LATIC anzuwenden, müssen Sie LATIC zunächst herunterladen. Öffnen Sie die Website <https://download.latic.software>. Über diesen Link werden Sie direkt zu der aktuellsten Version von LATIC auf der Plattform GitHub weitergeleitet. Unter dem Reiter „Assets“, können Sie den .zip Ordner für Ihr Betriebssystem herunterladen. Im Falle, dass Sie Windows nutzen, laden Sie den Ordner „LATIC_Windows.zip“ herunter. Im Falle, dass Sie Ubuntu oder Debian (Linux) nutzen, laden Sie den Ordner „LATIC_Ubuntu_Debian.zip“ herunter. Im Falle, dass Sie (1) macOS nutzen, (2) andere Linux Betriebssysteme als die oben genannten nutzen oder (3) LATIC über die Konsole bzw. das Terminal starten möchten, laden Sie den Ordner „LATIC_jar.zip“ herunter.

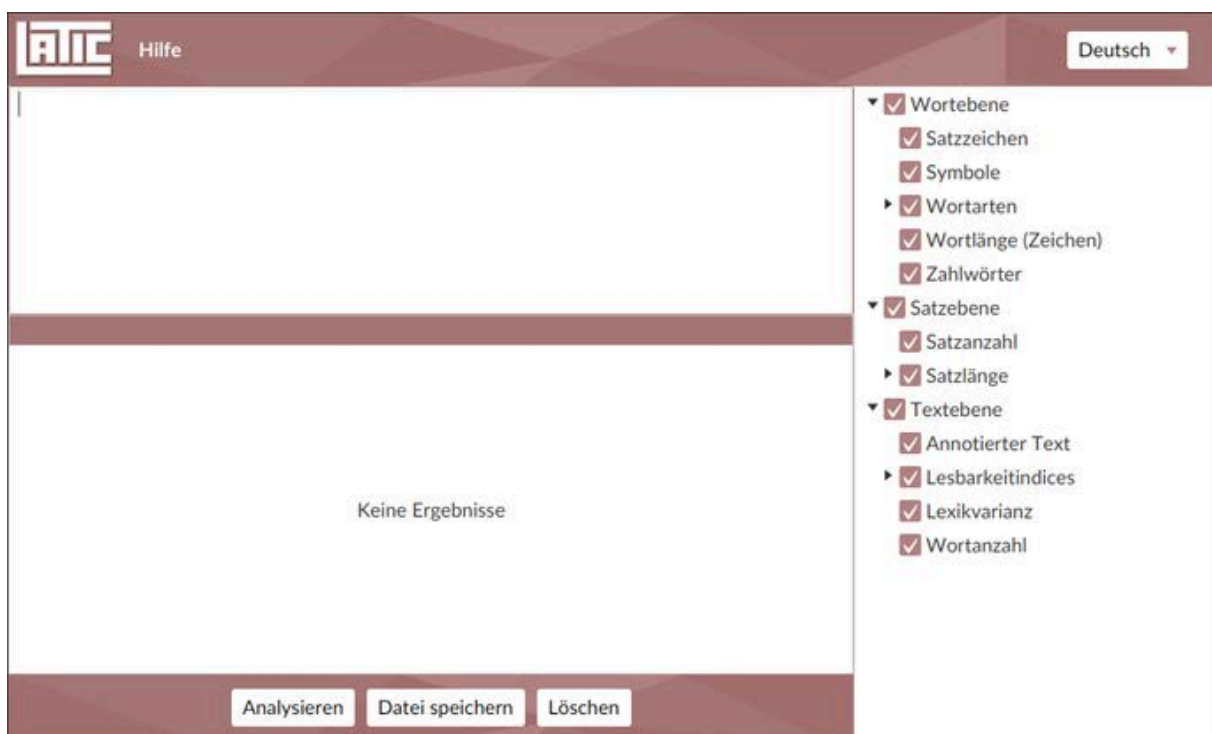
Klicken Sie mit einem Rechtsklick auf den .zip Ordner und entpacken Sie ihn. Öffnen Sie nun den Ordner. In dem Ordner finden Sie weitere Instruktionen wie Sie LATIC (installieren und) starten können.

4.2 Anwendung

Wenn Sie LATIC öffnen, sehen Sie die Bedienoberfläche (Abbildung 1). Um ein Item zu analysieren, führen Sie folgende Schritte durch:

Abbildung 1

Bedienoberfläche von LATIC



1. Stellen Sie zunächst oben rechts die Sprache des Items ein, das Sie analysieren möchten.
2. Geben Sie nun das zu analysierende Item in das obere Textfeld ein.
3. Wählen Sie rechts in der Spalte die Itemcharakteristika aus, die analysiert werden sollen. Sie können hier zwischen verschiedenen Itemcharakteristika der Wort-, Satz- und Textebene wählen.
4. Klicken Sie anschließend auf den Knopf „Analysieren“.

5. Die Ergebnisse der Analyse werden nun ausgegeben.
6. Um die Ergebnistabelle zu speichern, klicken Sie auf „Datei speichern“.
7. Um die Ergebnistabelle zu löschen, klicken Sie auf „Löschen“. Es erscheint ein Pop-up Fenster, das Sie bittet Ihre Entscheidung, das Ergebnis zu löschen, zu bestätigen.
8. Sollten Sie Hilfe benötigen, finden Sie unter dem Reiter „Hilfe“ mögliche Hilfestellungen.

4.3 Allgemeine Hinweise

Um zuverlässige Ergebnisse zu erhalten, ist es unerlässlich einige Faktoren zu berücksichtigen. Erstens sollten Sie auf eine korrekte Rechtschreibung inklusive der Groß- und Kleinschreibung achten. Ansonsten ergeben sich überdurchschnittlich häufig Fehler in den Analysen.

Zweitens empfehlen wir auf Abkürzungen zu verzichten. So erhalten Sie zuverlässige Ergebnisse in Bezug auf bestimmte Itemcharakteristika wie beispielsweise der durchschnittlichen Wort- und Satzlänge.

Drittens wird jede Zeichenkette als eine Einheit annotiert. Das bedeutet auch, dass Eigennamen, die aus mehreren Wörtern bestehen, mehrfach annotiert und gezählt werden. In dem Satz „Das *Rote Meer* hat einen hohen Salzgehalt.“ erhalten beide Wortteile des Roten Meeres jeweils das Tag PROP (Eigenname). Entsprechend zählt LATIC in diesem Satz zwei Eigennamen.

Viertens ist es LATIC (noch) nicht möglich, gendergerechte Sprache zu erkennen, die anhand von Satzzeichen getätigt wird (z. B. „Die *Schüler:innen* lernen.“, „Die *Lehrer/-innen* unterrichten.“). Wir empfehlen daher auf genderneutrale Sprache auszuweichen (z. B. „Die *Lehrkräfte* unterrichten.“). Im Falle, dass dies nicht möglich ist, empfehlen wir (1) die Satzzeichen für die Analyse ersatzlos zu streichen oder (2) die Paarform (z. B. „Die *Schülerinnen und Schüler* lernen.“) zu wählen, um zuverlässige Ergebnisse zu erhalten.

Fünftens muss zwischen einem Zahlwort und einer Maßeinheit ein Leerzeichen gesetzt werden (z. B. „Der Äquator ist mehr als *40.000 km* lang.“). Auf diese Weise werden sowohl die Zahlwörter als auch die Maßeinheiten als eigenständige Einheiten erkannt und erhalten jeweils ein eigenes Tag. Maßeinheiten mit Potenzzahlen (z. B. „Die Fläche ist *20 m²* groß.“) werden in der Regel nicht als eine Maßeinheit erkannt und sollten ausgeschrieben werden, um korrekt annotiert zu werden.

5 Funktionen von LATIC

Diese Funktionen gelten nur für die deutsche Sprache. Für die Funktionen in der englischen Sprache, schauen Sie bitte in die englische Dokumentation von LATIC.

5.1 Analyse auf Wortebene

Jede Zeichenkette erhält ein eigenes Tag. Das Annotieren basiert auf dem Stuttgart-Tübingen-TagSet (Schiller et al., 1999). LATIC ermöglicht die Zählung der verschiedenen Wortarten. Eine Übersicht aller Tags, die mithilfe des Stanford CoreNLP 4.2.1 (Manning et al., 2014) von LATIC ausgegeben werden, finden Sie in Tabelle 1.

5.1.1 Adjektive

Adjektive, auch Eigenschafts- oder Beiwörter genannt, sind Wörter, die ein Nomen modifizieren, indem sie dem Nomen bestimmte Eigenschaften oder Merkmale zuschreiben (Bibliographisches Institut, 2020). Bei der Annotation von Adjektiven ist zu beachten, dass ordinale Zahlen (z. B. „Das *dritte* Haus gehört meinem Onkel.“) ebenfalls als Adjektive gelistet werden. Adjektive erhalten das Tag ADJ (z. B. „Die Lampe leuchtet *hell*.“).

5.1.2 Adpositionen

Adpositionen gelten in der Linguistik als Überbegriff für Prä- und Postpositionen (Kurvon & Adler, 2008). Postpositionen bezeichnen dabei im Gegensatz zu Präpositionen Adpositionen, die hinter einem Wort stehen (z. B. „Meiner Meinung *nach* ist die Aussage richtig.“). Adpositionen erhalten das Tag ADP (z. B. „Die Kinder fahren *nach* Hause.“).

Tabelle 1*Übersicht aller Tags*

Tag	Beschreibung	Beispiel
ADJ	Adjektive	Viele Menschen finden die Weihnachtszeit <i>besinnlich</i> .
ADP	Adpositionen	<i>Auf</i> dem Bahnhof sieht man einige Touristen.
ADV	Adverbien	<i>Hier</i> können wir unser Zelt aufbauen.
AUX	Hilfsverben	Gestern <i>ist</i> eine Band im Stadtpark aufgetreten.
CCONJ	Koordinierende Konjunktionen	Zum Frühstück gibt es Brot <i>und</i> Marmelade.
DET	Determinanten	Kinder sind in <i>der</i> Auswahl von Kleidung wählerisch.
INTJ	Interjektionen	<i>Oje</i> , das tat bestimmt ganz schön weh!
NOUN	Nomen	Manchmal schneit es im <i>Winter</i> .
NUM	Zahlwörter	Das Grundgesetz für die Bundesrepublik Deutschland wurde <i>1949</i> erlassen.
PART	Partikel	Sie weiß <i>nicht</i> , ob sie es pünktlich zum Termin schafft.
PRON	Pronomen	<i>Er</i> fand den Vortrag sehr spannend.
PROPN	Eigennamen	<i>Deutschland</i> besteht aus 16 Bundesländern.
PUNCT	Satzzeichen	Gehört das Haus, das auf der anderen Straßenseite steht, deinen Großeltern?
SCONJ	Subordinierende Konjunktionen	<i>Obwohl</i> ich schlecht geschlafen habe, bin ich nicht müde.
SYM	Symbole	= + / \ ~ # > ^
VERB	Verben	Er <i>hat</i> seinen Arzttermin völlig <i>vergessen</i> .
X	Unbekannt oder ungewiss	

5.1.3 Adverbien

Adverbien beziehen sich auf Verben, Adjektive oder andere Adverbien. Sie definieren zusätzliche Angaben zu einem Ort (Lokaladverbien), einer Zeit (Temporaladverbien), einem Grund (Kausaladverbien) oder zu einer Art und Weise (Modaladverbien) (Eisenberg, 2013).

Adverbien erhalten das Tag ADV (z. B. „Ich tanze *gerne*.“).

5.1.4 Determinanten

Determinanten ersetzen Nominalgruppen oder bilden Teile dieser Nominalgruppen. Subklassen von Determinanten sind definierte Artikel, demonstrative, possessive, indefinite und interrogative Determinanten (Engel, 1979). Im Universal Dependencies Project (2014–2020) wird darauf hingewiesen, dass die Trennung von Determinanten und Pronomen (siehe Abschnitt 5.1.11) nicht immer eindeutig ist. Determinanten erhalten das Tag DET (z. B. „*Das* Gebäude ist im viktorianischen Stil gebaut.“).

5.1.5 Eigennamen

Eigennamen sind Namen von Dingen oder Individuen und gehören zur Wortart der Nomen (Eisenberg, 2013). Bekannte Akronyme von Eigennamen sollten ebenfalls als Eigennamen erkannt werden. Eigennamen erhalten das Tag PROP (z. B. „Die *UNO* ist ein Zusammenschluss aus 193 Staaten.“).

5.1.6 Hilfsverben

Hilfsverben sind Funktionswörter, die gemeinsam mit einem Verb zusammengesetzt werden, um bestimmte grammatische Merkmale auszudrücken (Hentschel & Weydt, 1990). Hilfsverben erhalten das Tag AUX (z. B. „Sie *sollte* ihn anrufen.“).

5.1.7 Interjektionen

Interjektionen sind Wörter, die wortähnliche Lautäußerungen darstellen. Mit ihnen sollen Gefühle, Aufforderungen oder bestimmte Laute zum Ausdruck gebracht werden (Bibliographisches Institut, 2020). LATIC annotiert nur primäre Interjektionen (z. B. „*Oh*

nein!“). Sekundäre Interjektionen werden nicht als solche erkannt (z. B. “*Donnerwetter!*“), sondern entsprechend ihrer ursprünglichen Wortart annotiert. Interjektionen erhalten das Tag INTJ (z. B. „*Wow*, du siehst toll aus!“).

5.1.8 Konjunktionen

Konjunktionen sind Wörter, die Satzgefüge miteinander verbinden (Bibliographisches Institut, 2020). Es kann zwischen koordinierenden sowie subordinierenden Konjunktionen unterschieden werden (Eisenberg, 2013).

5.1.8.1 Koordinierende Konjunktionen. Koordinierende Konjunktionen verbinden Sätze gleicher Form; beispielsweise einen Hauptsatz mit einem Hauptsatz oder einen Nebensatz mit einem Nebensatz (ebd.). Koordinierende Konjunktionen werden synonym auch als nebenordnende und beiordnende Konjunktionen bezeichnet. Sie erhalten das Tag CCONJ (z. B. „Die Bäckerin kaufte Mehl *und* Zucker ein.“).

5.1.8.2 Subordinierende Konjunktionen. Subordinierende Konjunktionen leiten Nebensätze ein (ebd.). Alternativ werden subordinierende Konjunktionen auch gehäuft als unterordnende Konjunktionen bezeichnet. Sie erhalten das Tag SCONJ (z. B. „Das Kind weinte, *weil* es so sehr lachte.“).

5.1.9 Nomen

Nomen, auch Hauptwörter oder Substantive genannt, bezeichnen Gattungen, Stoffe sowie Kollektiva mit spezifischen Eigenschaften (Eisenberg, 2013). Eigennamen werden gesondert annotiert (siehe Abschnitt 5.1.5). Nomen erhalten das Tag NOUN (z. B. „*Sauerstoff* ist ein chemisches *Element*.“).

5.1.10 Partikel

Partikel sind Funktionswörter, die nicht flektierbar sind und nicht den Wortarten der Präpositionen, Adverbien oder Konjunktionen zugeordnet werden können (Eisenberg, 2013). Sogenannte verbale Partikel (z. B. “to give *up*“), wie sie in der englischen Sprache genannt

werden, werden im Deutschen als Adpositionen (z. B. “Ich gebe *nach*.”) bzw. Adverbien annotiert (z. B. “Ich kaufe *ein*.”). Partikel erhalten das Tag PART (z. B. „Ich hoffe, meinen Vortrag gut *zu* meistern.“).

5.1.11 Pronomen

Pronomen ersetzen Nomen oder Nominalphrasen (Eisenberg, 2013). Im Deutschen ist es dem Stanford CoreNLP 4.2.1 (Manning et al., 2014) nicht möglich, unterschiedliche Arten von Pronomen zu unterscheiden. Daher wird in LATIC nur die Anzahl aller Pronomen ausgegeben. Sie erhalten das Tag PRON (z. B. „*Wir* essen nur Gemüse, *welches* biologisch angebaut wurde.“).

5.1.12 Satzzeichen

Satzzeichen verdeutlichen die syntaktische Struktur eines Textes (Universal Dependencies Project, 2014–2020). Sie erhalten das Tag PUNCT (z. B. „Das Haus brennt! Ruft die Feuerwehr!“).

5.1.13 Symbole

Symbole unterscheiden sich von Worten hinsichtlich ihrer Form und/ oder ihrer Funktion (Universal Dependencies Project, 2014–2020). Sie erhalten das Tag SYM (z. B. „Die Anwältin verweist ihren Kollegen auf § 2.“).

5.1.14 Unbekannt oder Ungewiss

Wortarten oder Zeichenketten, die LATIC mithilfe des Stanford CoreNLP 4.2.1 (Manning et al., 2014) keinem Tag zuordnen kann, werden als unbekannt bzw. ungewiss annotiert. Diese Wörter oder Zeichenketten erhalten das Tag X.

5.1.15 Verben

Ein Verb gibt ein Geschehen, eine Tätigkeit oder einen Vorgang wieder (Bibliographisches Institut, 2020). Im Rahmen des Stanford CoreNLP 4.2.1 (Manning et al., 2014), werden Hilfsverben gesondert annotiert (siehe Abschnitt 5.1.6). Partizipien werden je

nach Kontext als Verben (z. B. „Ich habe *gegessen*.“) oder als Adjektive (z. B. „Das Gerät ist zum Gebrauch *geeignet*.“) annotiert. Nominalisierte Verben (z. B. „Das *Laufen* macht Spaß.“) sollten als Nomen annotiert werden. Verben erhalten das Tag VERB (z. B. „Die Katze *schläft* auf dem Bett.“).

5.1.16 Wortlänge

Die durchschnittliche Wortlänge wird berechnet, indem die Zeichenanzahl (ausgenommen von Satz- und Leerzeichen) durch die Wortanzahl dividiert wird. Tabelle 2 beinhaltet zwei Beispiele zur Veranschaulichung.

Tabelle 2

Beispiele zur Analyse der durchschnittlichen Wortlänge

Beispiele				Durchschnittliche Wortlänge
<u>Katzen</u> 6 Zeichen	<u>haben</u> 5 Zeichen	<u>vier</u> 4 Zeichen	<u>Beine.</u> 5 Zeichen	5,0 Zeichen
<u>Katzen</u> 6 Zeichen	<u>haben</u> 5 Zeichen	<u>4</u> 1 Zeichen	<u>Beine.</u> 5 Zeichen	4,25 Zeichen

5.1.17 Zahlwörter

Zahlwörter fungieren laut dem Universal Dependencies Project (2014–2020) meistens als Determinanten, Adjektive oder Pronomen. Dabei werden Zahlwörter als Zahlen und als Wörter definiert, die eine Beziehung zu Zahlen herstellen. Um zuverlässige Ergebnisse zu erhalten, sollten die Zahlen in der arabischen Form geschrieben werden. Römische sowie ausgeschriebene Zahlen werden nicht immer korrekt vom Stanford CoreNLP 4.2.1 (Manning et al., 2014) als Zahlwörter erkannt. Non-kardinale Zahlwörter (z. B. „*erstens*“, „*doppelt*“) werden als Adjektive oder Adverbien annotiert. Zahlwörter (inklusive Dezimalzahlen) erhalten das Tag NUM (z. B. „Ein Schaltjahr hat *366* Tage.“).

5.2 Analyse auf Satzebene

5.2.1 Satzanzahl

Zeichenketten sind Sätze, wenn sie durch die folgenden Satzzeichen beendet werden: Punkte (.), Fragezeichen (?) sowie Ausrufezeichen (!). Zeilenumbrüche, Doppelpunkte (:) sowie Semikola (;) beenden keine Sätze. Wenn Sätze mit Doppelpunkten oder Semikola als eigenständige Sätze gezählt werden sollen, empfehlen wir, diese durch Punkte zu ersetzen. Tabelle 3 beinhaltet zwei Beispiele zur Veranschaulichung.

Tabelle 3

Beispiele zur Analyse der Satzanzahl

Beispiele	Satzanzahl
LATIC ist ein Analysetool. Es steht für “Linguistisches Analysetool für Itemcharakteristika”.	2
LATIC ist ein Analysetool: Es steht für “Linguistisches Analysetool für Itemcharakteristika”.	1

5.2.2 Satzlänge

Die durchschnittliche Satzlänge kann jeweils anhand (1) der Zeichenanzahl inklusive Leerzeichen, (2) der Zeichenanzahl ohne Leerzeichen sowie (3) der Wortanzahl (inkl. Symbolen und Zahlen) ausgegeben werden. Dabei wird die Zeichen- bzw. Wortanzahl jeweils durch die Satzanzahl dividiert. Leer- sowie Satzzeichen werden bei der Berechnung der Satzlänge nicht berücksichtigt. Tabelle 4 beinhaltet Beispiele zur Veranschaulichung.

Tabelle 4

Beispiele zur Analyse der durchschnittlichen Satzlänge

Beispiele	Durchschnittliche Satzlänge
<u>Morgen</u> 6 Zeichen	44,0 Zeichen (mit Leerzeichen) 38,0 Zeichen (ohne Leerzeichen)
<u>findet</u> 6 Zeichen	
<u>im</u> 2 Zeichen	7,0 Wörter
<u>Hörsaal</u> 7 Zeichen	
<u>eine</u> 4 Zeichen	23,5 Zeichen (mit Leerzeichen) 20,0 Zeichen (ohne Leerzeichen)
<u>Klausur</u> 7 Zeichen	
<u>statt.</u> 5 Zeichen	4,5 Wörter
<u>Mein</u> 4 Zeichen	
<u>Freund</u> 6 Zeichen	
<u>hat</u> 3 Zeichen	
<u>eine</u> 4 Zeichen	
<u>Hündin.</u> 6 Zeichen	
<u>Ihr</u> 3 Zeichen	
<u>Name</u> 4 Zeichen	
<u>ist</u> 3 Zeichen	
<u>Bella.</u> 5 Zeichen	

5.3 Analyse auf Textebene

5.3.1 Annotiertes Item

Mit der Funktion des annotierten Items gibt LATIC die Tags aus, die jeder einzelnen Zeichenkette vergeben wird. Leerzeichen erhalten dabei kein eigenes Tag. Abbildung 2 beinhaltet ein Beispiel zur Veranschaulichung.

Abbildung 2

Beispiel eines annotierten Items

<u>Säugetiere</u>	<u>gehören</u>	<u>zur</u>	<u>Klasse</u>	<u>der</u>	<u>Wirbeltiere</u>	<u>und</u>	<u>zeichnen</u>	<u>sich</u>	<u>durch</u>	<u>bestimmte</u>
NOUN	VERB	ADV	NOUN	DET	NOUN	CCONJ	VERB	PRON	ADP	ADJ

<u>Charakteristika</u>	<u>aus</u>	<u>.</u>	<u>Dazu</u>	<u>gehört</u>	<u>zum</u>	<u>Beispiel</u>	<u>.</u>	<u>dass</u>	<u>die</u>	<u>meisten</u>	<u>Säugetiere</u>
NOUN	ADP	PUNCT	ADV	VERB	ADP	NOUN	PUNCT	SCONJ	DET	ADJ	NOUN

<u>ihre</u>	<u>Jungen</u>	<u>lebend</u>	<u>gebären</u>	<u>.</u>
DET	NOUN	ADJ	VERB	PUNCT

5.3.2 Lesbarkeitsindices

Zurzeit ist es möglich, mithilfe von LATIC den Lesbarkeitsindex LIX zu berechnen. Weitere Lesbarkeitsindices sollen in Zukunft in LATIC implementiert werden.

5.3.2.1 LIX. Der Lesbarkeitsindex, abgekürzt LIX, wurde ursprünglich von Carl-Hugo Björnsson (1968) entwickelt. Zur Bestimmung der Lesbarkeit eines Items werden beim LIX

die Satzanzahl, die Wortanzahl sowie die Anzahl an langen Wörtern (definiert als Wörter mit mehr als sechs Buchstaben) herangezogen.

Die Formel für die Berechnung des LIX lautet:

$$LIX = \frac{Wortanzahl}{Satzanzahl} + \frac{\text{lange Wörter}}{Wortanzahl} * 100$$

Der LIX kann Werte zwischen 0 und 100 annehmen. Anhand der Werte kann eine Einschätzung zur Schwierigkeit eines Items vorgenommen werden (siehe Tabelle 5; zitiert nach Lenhard, W. & Lenhard, A., 2014–2017). Tabelle 6 beinhaltet zwei Beispiele zur Veranschaulichung.

Tabelle 5

Interpretation der LIX Werte

Itemschwierigkeit	LIX Werte	Ähnlichkeit mit
Sehr leicht	bis 29	
Leicht	30–39	Kinder- und Jugendliteratur
Mittel	40–49	Belletristik
Schwierig	50–59	Sachliteratur
Sehr schwierig	über 60	Fachliteratur

Tabelle 6

Beispiele zur Analyse des LIX

Beispiele	LIX Wert	Itemschwierigkeit
Mein Name ist Tim. Ich habe einen Bruder und eine Schwester. Sie heißen Leon und Susi. Wir haben auch einen Hund. Sein Name ist Bello und ich mag ihn sehr gerne.	9,43	Sehr leicht

Beispiele	LIX Wert	Itemschwierigkeit
Desoxyribonukleinsäure ist eine aus unterschiedlichen Desoxyribonukleotiden aufgebaute Nukleinsäure. Sie trägt die Erbinformation bei allen Lebewesen und vielen Viren. Das langkettige Polynukleotid enthält in Abschnitten von Genen besondere Abfolgen seiner Nukleotide.	56,67	Schwierig

Anmerkung. Das zweite Beispiel stammt von Wikipedia („Desoxyribonukleinsäure“, 2021).

5.3.3 Lexikvarianz

Die Lexikvarianz, auch Type-Token-Relation genannt, stellt die Anzahl verschiedener Wörter dividiert durch die Wortanzahl dar (Johansson, 2009). Tabelle 7 beinhaltet Beispiele zur Veranschaulichung.

Tabelle 7

Beispiele zur Analyse der Lexikvarianz

Beispiele	Lexikvarianz
<i>Eine Firma wurde beauftragt einen Swimmingpool zu bauen. Normalerweise brauchen drei Mitarbeiter dafür drei Tage. Wie viele Tage würde ein Mitarbeiter brauchen, wenn er den Swimmingpool alleine bauen müsste?</i>	Anzahl verschiedener Wörter: 23 Gesamte Wortanzahl: 29 Lexikvarianz: 0,79
<i>Fischer Fritz fischt frische Fische. Frische Fische fischt Fischer Fritz.</i>	Anzahl verschiedener Wörter: 5 Gesamte Wortanzahl: 10 Lexikvarianz: 0,5

Anmerkung. Die Wörter, die zur Anzahl der verschiedenen Wörter gezählt werden, sind kursiv gedruckt.

5.3.4 Wortanzahl

Zeichenketten, die durch Leer- und/oder Satzzeichen voneinander getrennt werden, werden als Wörter erkannt. Symbole und Zahlwörter werden ebenfalls als Wörter erkannt.

Tabelle 8 beinhaltet Beispiele zur Veranschaulichung.

Tabelle 8

Beispiele zur Analyse der Wortanzahl

Beispiele	Wortanzahl
Es war einmal ein Junge. Er lebte mit seinen Eltern im Walde. Eines Tages beobachtete der Junge ein Eichhörnchen, das gerade eine Nuss vergrub. „Das Eichhörnchen macht sich bestimmt bereit für den Winterschlaf“, dachte der Junge.	36
Pinguine sind Vögel, die nicht fliegen können. Dabei gibt es viele verschiedene Pinguinarten. Ein Kaiserpinguin wird zum Beispiel bis zu 1,20 Meter groß und kann bis zu 40 Kilogramm wiegen.	30

6 Evaluation

Zur Evaluation von LATIC unter der Verwendung des Stanford CoreNLP 4.2.1 (Manning et al., 2014) wurde der TIGER Corpus 2.2 verwendet (Brants et al., 2004). Der TIGER Corpus besteht aus etwa 50.000 Sätzen bzw. 900.000 Tokens, die aus der deutschen Zeitung „Frankfurter Rundschau“ entnommen wurden. Die Tokens wurden semi-automatisch annotiert (Brants & Plaehn, 2000). Dabei wurde auf das TIGER Treebank Tagset (Smith, 2003) als auch das Penn Treebank Tagset (Santorini, 1990) zurückgegriffen.

Zunächst wurden zur Evaluation 10.002 Tokens des TIGER Corpus 2.2 (Brants et al., 2004) übernommen. Da bereits während des Testens deutlich wurde, dass die korrekte Rechtschreibung ausschlaggebend für eine korrekte Annotation ist, wurden fünf Tokens

entfernt (z. B. Wörter, die ausschließlich in Großbuchstaben verfasst wurden). Final wurden entsprechend $N = 9.997$ Tags miteinander verglichen.

Im nächsten Schritt wurden die Tagsets überarbeitet, um die einzelnen Tokens vergleichen zu können. Dafür mussten die Tagsets des TIGER Corpus 2.2 (Brants et al., 2004) sowie des Stanford CoreNLP 4.2.1 (Manning et al., 2014) ineinander überführt werden. Beispielsweise annotiert das Stanford CoreNLP 4.2.1 (ebd.) in der deutschen Sprache Hilfsverben mithilfe des Tags AUX. Innerhalb des TIGER Corpus 2.2 (Brants et al., 2004) werden Hilfsverben jedoch anhand von mehreren Tags deutlich differenzierter annotiert. Nach dieser Überführung der Tagsets stimmten die Annotationen des TIGER Corpus 2.2 (ebd.) und die Annotationen des Stanford CoreNLP 4.2.1 (Manning et al., 2014) in 90,96 % Fällen überein ($n = 9.093$).

An dieser Stelle sei anzumerken, dass der Abgleich der Annotationen korrigiert werden musste. Erstens waren einige Tags, die im TIGER Corpus 2.2 (Brants et al., 2004) verwendet wurden, nicht in das Tagset des Stanford CoreNLP 4.2.1 (Manning et al., 2014) überführbar. Dazu gehörten unter anderem die Tags TRUNC (abgetrenntes Kompositionserstglieder) und KOKOM (Vergleichskonjunktionen). Diese wurden daher aus dem Abgleich gestrichen. Zweitens wurden einige Tags in den Tagsets unterschiedlich definiert. Dazu gehörten unter anderem Partikel sowie Eigennamen. Während verbale Partikel im TIGER Corpus 2.2 (Brants et al., 2004) als solche annotiert werden, werden diese anhand des Stanford CoreNLP 4.2.1 (Manning et al., 2014) bewusst als Adpositionen oder Adverbien annotiert (siehe Abschnitt 5.1.10). Drittens wird im Stanford CoreNLP 4.2.1 (ebd.) jede Zeichenkette separat annotiert. Beispielsweise erhält das Wort „Informatik-Dienstleistungsunternehmen“ vom Stanford CoreNLP 4.2.1 (ebd.) die Tags Nomen – Satzzeichen – Nomen, während es im TIGER Corpus 2.2 (Brants et al., 2004) als Nomen

annotiert wurde. Nach dieser sehr konservativen Korrektur waren 93,85 % der Annotationen des Stanford CoreNLP 4.2.1 (Manning et al., 2014) korrekt ($n = 9.271$).

Anhand der ersten 5.000 Annotationen wurden die häufigsten Fehlerquellen untersucht, die dem Stanford CoreNLP 4.2.1 (Manning et al., 2014) unterlaufen. Die meisten Fehler beim Annotieren wurden bei der Unterscheidung von Determinanten und Pronomen gemacht (ca. 21,74 % aller Fehler). Dies war jedoch zu erwarten, da die Unterscheidung zwischen Pronomen und Determinanten nicht immer eindeutig ist (Universal Dependencies Project, 2014–2020). Die zweithäufigste Fehlerquelle war die Unterscheidung zwischen Nomen und Eigennamen (ca. 13,71 % aller Fehler). Die dritthäufigste Fehlerquelle war die Unterscheidung von Adjektiven und Adverbien (ca. 9,36 % aller Fehler).

Die Korrektheit der Annotationen variiert innerhalb und zwischen verschiedenen Tagger-Softwares. Diese Varianz kommt dabei unter anderem durch die verschiedenen Corpora und Textgenres zustande, die zur Evaluation der Annotationen verwendet werden. Beispielsweise wird eine Korrektheit der Annotationen von nur 81,16 % bei spontansprachlichen Daten erreicht (z. B. Westpfahl & Schmidt, 2013), während der prozentuale Anteil korrekter Annotationen bei Zeitungsartikeln und erklärenden Texten mit bis zu 98,25 % deutlich höher liegt (Giesbrecht & Evert, 2009). In Anlehnung an vorheriger Forschung erzielen die Annotationen durch das in LATIC implementierte Stanford CoreNLP 4.2.1 (Manning et al., 2014) gute Ergebnisse.

7 Literatur

- Björnsson, C. H. (1968). *Læsbarhed* [Lesbarkeit]. Gad.
- Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G., & Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Journal of Language and Computation*, 2(4), 597–620. <https://doi.org/10.1007/s11168-004-7431-3>
- Brants, T., & Plaehn, O. (2000). Interactive Corpus Annotation. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, & G. Steinhauer (Hrsg.), *Proceedings of the second international conference on language resources and evaluation (LREC-2000)*. European Language Resources Association.
- Chowdhury, G. G. (2005). Natural language processing. *Annual Review of Information Science and Technology*, 37(1), 51–89. <https://doi.org/10.1002/aris.1440370103>
- Cruz Neri, N., Guill, K., & Retelsdorf, J. (2021). Language in science performance: Do good readers perform better? *European Journal of Psychology of Education*, 36(1), 45–61. <https://doi.org/10.1007/s10212-019-00453-5>
- Bibliographisches Institut (2020). *Duden Wörterbuch*. <https://www.duden.de/>
- Desoxyribonukleinsäure (2021, Januar 25). In *Wikipedia*. <https://de.wikipedia.org/wiki/Desoxyribonukleins%C3%A4ure>
- Eisenberg, P. (2013). *Grundriss der deutschen Grammatik: Band 2: Der Satz* (4. Auflage). Springer.
- Engel, U. (1979). Syntaktische Strukturen. In H. Steger (Hrsg.), *Das Zertifikat Deutsch als Fremdsprache* (S. 67–119). Deutscher Volkshochschul-Verband e.V. Bonn.
- Giesbrecht, E., & Evert, S. (2009). Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In I. Alegria, I. Leturia & S. Sharoff

- (Hrsg.), *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, San Sebastian, Spain. http://www.stefan-evert.de/PUB/GiesbrechtEvert2009_Tagging.pdf
- Hentschel, E., & Weydt, H. (1990). *Handbuch der deutschen Grammatik*. De Gruyter.
- Heppt, B., Haag, N., Böhme, K., & Stanat, P. (2015). The role of academic-language features for reading comprehension of language-minority students and students from low-SES families. *Reading Research Quarterly*, 50(1), 61–82. <https://doi.org/10.1002/rrq.83>
- Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working Papers*, 53(2008), 61–79.
- Kurvon, D., & Adler, S. (2008). Introduction. In D. Kurzon & S. Adler (Hrsg.), *Adpositions. Pragmatic, semantic and syntactic perspectives* (S. 1–12). John Benjamins B.V.
- Lenhard, W. & Lenhard, A. (2014–2017). *Berechnung des Lesbarkeitsindex LIX nach Björnson*. <http://www.psychometrica.de/lix.html>.
- Manning, C. D., Surdeani, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014, June 23-24). *The Stanford CoreNLP Natural Language Processing Toolkit*. Proceedings of 52nd Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, Maryland, United States. <https://www.aclweb.org/anthology/P14-5010.pdf>
- Santorini, B. (1990). *Part-of-speech tagging guidelines for the Penn Treebank Project*. Department of Computer and Information Science, University of Pennsylvania. <https://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf>
- Schiller, A., Teufel, S., & Stöckert, C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Institut für maschinelle Sprachverarbeitung. <https://www.ims.uni-stuttgart.de/documents/ressourcen/lexika/tagsets/stts-1999.pdf>

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105–126.
https://doi.org/10.1207/s15326977ea1102_2

Smith, G. (2003). *A brief introduction to the TIGER treebank, version 1*. Universität Stuttgart.
https://www.ims.uni-stuttgart.de/documents/ressourcen/korpora/tiger-corpus/annotation/tiger_introduction.pdf

Universal Dependencies (2014–2020). *Universal POS tags*.
<https://universaldependencies.org/u/pos/>

Westpfahl, S., & Schmidt, T. (2013). POS für(s) FOLK – Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. *Journal for Language Technology and Computational Linguistics, 1*, 139–156.

White, S. (2012). Mining the text: 34 text features that can ease or obstruct text comprehension and use. *Literacy Research and Instruction, 51*(2), 143–164.
<https://doi.org/10.1080/19388071.2011.553023>