Institut für Informatik
Goethe-Universität Frankfurt

Bachelor Thesis

Optimierung beim Annotieren mehrsprachiger Münzdatensätze im Kontext des Natural Language Processing

Katrin Peikert

30.03.2022

Betreut von
Dr. Karsten Tolle

Databases and Information Systems (DBIS)

Inhaltsverzeichnis

1 Motivation **1**

2 Grundlagen **3**
2.1 Beschreibung des *Corpus Nummorum Online* 3
2.2 Die grundlegenden Arbeiten von Klinger, 2018 und Deligio und Gencer, 2021 . 4
2.3 Flektion im Deutschen und Englischen 6
2.3.1 Deutsche Flektion 6
2.3.2 Englische Flektion 8
2.4 Einführung in Lemmatisierung und Stemming 9
2.4.1 Lemmatisierung 10
2.4.2 Stemming 11
2.5 Verwendete Metriken 12

3 Modellierung **13**
3.1 Problembeschreibung 13
3.2 Programmbeschreibung 15

4 Durchführung und Ergebnisse **20**
4.1 Durchführung 20
4.1.1 Analyse der deutschen Annotationen 21
4.1.2 Analyse der englischen Annotationen 23
4.2 Besprechung der nicht entdeckten Entitäten 26
4.3 Besprechung der neu entdeckten Entitäten 27
4.4 Vergleich mit anderen Lemmatisierungs- und Stemmingalgorithmen 29

5 Fazit und Ausblick **33**

A Tabellarische Aufzählung der annotierten Entitäten **35**
A.1 Übersicht der gefundenen Entitäten der Kombi-Annotation (Englisch) 35
A.2 Übersicht der gefundenen Entitäten der Kombi-Annotation (Deutsch) 36
A.3 Übersicht der nur von der manuellen Annotation gefundenen Entitäten (Deutsch) 37
A.4 Übersicht der nur von der manuellen Annotation gefundenen Entitäten (Englisch) 37
A.5 Vergleich der alternativen Lemmatisierungs- und Stemmingalgorithmen der englischen Annotation 38
A.6 Vergleich der alternativen Lemmatisierungs- und Stemmingalgorithmen der deutschen Annotation 39

B Abbildungsverzeichnis **41**

C Literatur **42**

1 Motivation

Vor fast 3000 Jahren wandelte sich das Handelssystem in Europa vom Gütertauschhandel zu einem Währungssystem mit Edelmetallen und später mit geprägten Edelmetallstücken, welche wir als Münzen bezeichnen (Davies (2002, S. 61–65)). Besonders interessant ist der Umstand, dass dabei Münzen in ihrer Prägung stets die sie umgebende Kultur widerspiegeln. In dem Verzeichnis *Corpus Nummorum Online* werden Münzen einer bestimmten Zeitperiode der Menschheitsgeschichte zu Forschungszwecken und Erhalt der kulturellen Geschichte katalogisiert. Interessierte Menschen werden somit in die Lage versetzt, sich einen Einblick in diesen Teil der Kulturgeschichte der Menschheit zu verschaffen und dabei die vielen Gemeinsamkeiten, aber auch Unterschiede der Münzen zu entdecken.

Corpus Nummorum Online enthält Einträge zu 22808 Münzen[1] aus Gebieten des antiken Griechenland, namens Moesia inferior, Thrakien, Mysien und der Troas. Für jede dieser Münzen sind potentiell viele verschiedene Informationen wie Alter, Herkunft, Gewicht, Größe, sowie detaillierte Beschreibungen der Prägungen angegeben. Um diese große Menge an Daten effizient verarbeiten zu können, ist der Einsatz computergestützter Methoden unabdingbar. In den Arbeiten von Klinger (2018) und Deligio und Gencer (2021) werden durch den Einsatz von Methoden des *Natural Language Processing* (NLP) diese Münzbeschreibungen analysiert, um darauf abgebildete Personen, Gegenstände, Tiere und Pflanzen zu erkennen, sowie die Verbindungen zwischen diesen Entitäten aufzudecken. Dadurch kann ermittelt werden, welche Zusammenhänge zwischen den Münzen des *Corpus Nummorum Online* existieren und welche historischen und kulturellen Ereignisse die Münzen und ihre Abbilder geprägt haben.

Die Analyse der Münzbeschreibungen ist in zwei Abschnitte gegliedert. Zuerst wird mithilfe eines auf *named entity extraction* trainierten Modells erkannt, welche Entitäten, d.h. welche Personen, Objekte, Pflanzen und Tiere, in den Münzbeschreibungen erwähnt werden. Im zweiten Teil wird die Relation dieser Entitäten zueinander identifiziert mithilfe der sog. *relation extraction*. Dabei wird ermittelt, wie die Entitäten innerhalb einer Münzbeschreibung miteinander verknüpft sind. So wird die Entität „Keule “ von der Entität „Herakles“’ „gehalten“. Deligio und Gencer (2021) hat zusätzlich identifiziert, welche Beziehungen von Entitäten zu einem Verb ohne Objekt bestehen.
Bei der Vorbereitung der vorliegenden Münzbeschreibungen sind allerdings Schwierigkeiten aufgetreten. Die Entitäten, die in *Corpus Nummorum Online* auftreten, liegen als Einträge der sog. Entitätenlisten vor. Die Wörter in der Entitätenlisten werden in den Münzbeschreibungen allerdings nur in der dort vorliegenden Wortform gesucht. Viele Entitäten werden deshalb nicht in den Münzbeschreibungen entdeckt, da diese in einer anderen Wortform vorliegen wie z.B. in Pluralform. Wenn beispielsweise in der Entitätenliste der Tiere der Eintrag *„Löwe“* steht, dann wird auch nur diese Wortform gefunden. Die Pluralform *„Löwen“* wird nicht erkannt werden.

[1] https://www.corpus-nummorum.eu/search/coins?q=(15.03.2022)

Deligio und Gencer (2021) löst dieses Problem durch manuelles Erweitern der Entitätenlisten mit allen Wortformen einer Entität. Dies benötigt allerdings zusätzlichen Aufwand, und es ist nicht garantiert, dass alle zusätzlichen Wortformen hinzugefügt werden. Aus diesen Problemen leitet sich das Thema dieser Arbeit her.

Zielsetzung

Ziel dieser Arbeit ist das automatische Erkennen von Entitäten der Münzbeschreibungen des *Corpus Nummorum Online* (CNO). Dies stellt einen alternativen ersten Verarbeitungsschritt in der in Deligio und Gencer (2021) vorgestellten Pipeline dar. Lediglich die Annotation zu Beginn des *NER*- Prozesses soll angepasst werden, und der Rest des Modells bleibt unverändert.
Verschiedene flektierte Formen einer Entität sollen sowohl im Deutschen, als auch im Englischen, auf dieselbe Grundform zurückgeführt werden. Die Formen werden durch die NLP-Methoden Lemmatisierung und Stemming auf diese Grundform reduziert. Es wird geprüft, welche der beiden Methoden die besseren Ergebnisse erzielt. Zusätzlich wird auch die Kombination beider Methoden erstellt und bewertet. Bei der Bewertung dessen müssen die erstellten Annotationen auf Fehler und neu entdeckte Entitäten manuell geprüft werden. Zusätzlich sollen Zusammenfügungen von Nomen, die Entitäten enthalten können, (Komposita) in ihre Teilworte zerlegt und potenzielle Entitäten erkannt werden. Dem Gesamtwort wird anschließend der korrekte Entitätstyp seines Teilwortes zugewiesen.
Das Programm wird mit Python und der Library SpaCy implementiert, welche bereits in Klinger (2018) und Deligio und Gencer (2021) verwendet wurden.

Aufbau

In *Abschnitt 2* werden die Daten des *Corpus Nummorum Online* vorgestellt, und die Arbeit von Klinger (2018) und Deligio und Gencer (2021) besprochen, besonders in Hinsicht auf Schwierigkeiten bei der Annotation der deutschen und englischen Daten und der Datenpflege. Darauf aufbauend werden die Schwierigkeiten der deutschen und englischen Flektion erläutert. Anschließend werden die Begriffe Lemmatisierung und Stemming definiert und Algorithmen dazu vorgestellt. In *Abschnitt 3* werden auf die Probleme für Lemmatisierung und Stemming in den CNO- Daten eingegangen und eine Lösung vorgestellt. Im *Abschnitt 4* wird dann die damit gewonnene Annotation evaluiert und mit der Annotation von Deligio und Gencer (2021) verglichen. Vorkommende Fehler werden kategorisiert und deren Ursprünge untersucht. Anschließend wird diese Annotation mit Annotationen durch andere Lemmatisierungs- und Stemmingalgorithmen verglichen, und gezeigt, inwiefern diese das Ergebnis verbessern oder verschlechtern. Abschließend wird in *Abschnitt 5* zusammengefasst, welche Erfolge das hier vorgestellte Programm bei der Annotation der CNO-Daten erzielt und ein Ausblick auf mögliche, zukünftige Verbesserungen gegeben.

2 Grundlagen

Diese Kapitel stellt die verwendete Münzdatenbank von *Corpus Nummorum Online* und die Forschungsergebnisse von Klinger (2018) und Deligio und Gencer (2021) vor. Hierbei wird besonders auf die Schwierigkeiten eingegangen, die Deligio und Gencer (2021) in ihrer Weiterentwicklung und Anpassung des Modells an das Deutsche entdeckt haben. Dann werden die Schwierigkeiten, die durch Flektion im Deutschen und Englischen auftreten besprochen und in Zusammenhang mit den Wortformen des *Corpus Nummorum* gestellt. Im nächsten Schritt werden Algorithmen für Lemmatisierung und Stemming vorgestellt und Bewertungsmetriken aufgezeigt.

2.1 Beschreibung des *Corpus Nummorum Online*

Die *Corpus Nummorum Online* Datenbank enthält die Daten, die im Laufe dieser Arbeit verarbeitet werden. Diese Datenbank ist als Webportal verfügbar und beinhaltet Informationen zu antiken Münzen aus Moesien, Thrakien, Mysien und Troas[2]. Dabei wird ein besonderer Fokus auf das Herkunftsgebiet und deren angenommene Prägestätten gelegt. Begonnen hat die Sammlung mit den Münzen aus dem Münzkabinett Berlin, welches Münzen aus 104 Prägestätten enthält, sowie Abdrücken von Münzen aus weiteren Sammlungen, welche in der Berlin-brandenburgischen Akademie der Wissenschaften aufbewahrt sind. Insgesamt sind heute Münzsammlungen aus 25 Ländern eingebunden.

Corpus Nummorum Online bietet Nutzern verschiedene Optionen zum Durchsuchen der Daten. Es kann nach vielen verschiedenen Kriterien gefiltert werden wie Prägestätten, Typ, Epoche und auch Gewicht und Materialien. Ebenfalls werden standardisierte Beschreibungen der Abbilder der Münzen sowohl auf Deutsch, als auch auf Englisch zur Verfügung gestellt. Diese Beschreibungen sind innerhalb dieser Arbeit von besonderem Interesse. Das folgende Beispiel zeigt, wie eine solche standardisierte Beschreibung in Englisch und im Deutschen aussehen kann:

Athena standing facing, head left, wearing long garment and helmet; holding patera in outstretched right hand and inverted spear in left arm; at her feet, shield placed on ground. (ID:191)

Athena stehend von vorn, Kopf nach links, im langen Gewand und mit Helm; in der vorgestreckten Rechten Patera, in der Linken nach unten gerichteten Speer haltend; vor ihr, Schild.(ID:191)

Der Standard[3] beinhaltet Regeln, die von jeder Beschreibung eines Münzabbildes eingehalten werden sollen. Er beinhaltet Vorschriften zu dem Vokabular, welches verwendet werden soll. Lateinische Begriffe sollen hauptsächlich verwendet werden (z.B. „Dionysus“ anstatt „Dionysos“). In englischen Phrasen soll auf einen Bindestrich verzichtet werden („lion skin“ anstatt

[2] https://www.corpus-nummorum.eu/about (19.1.2022)
[3] https://www.corpus-nummorum.eu/pdf/ExternalCoinEntry.pdf (20.1.2022)

„lion-skin“). Ebenfalls ist eine Beschreibungsreihenfolge für das Münzmotiv festgelegt. Zuerst soll die abgebildete Person identifiziert werden, dann deren Kleidung und Haltung beschrieben werden, und danach die Objekte in der rechten Hand gefolgt von den Objekten in der linken Hand. Die einzelnen Satzbestandteile Person, Kleidung und Körperorientierung sollen von Kommata getrennt werden. Enthält ein Motiv mehrere Figuren werden zuerst die benannten Personen genannt und danach die Unbenannten,die mit einem Semikolon voneinander getrennt werden. Architektur soll detailliert beschrieben werden, wohingegen die Beschreibung von Kult-Statuen kurz zu fassen ist.

Die Daten von *Corpus Nummorum Online* liegen als SQL-Datenbankauszug vor, welcher 5514 verschiedene Münzbeschreibungen enthält, welche in dieser Arbeit als *Designs* bezeichnet werden. Aus der Datenbank können für ein Münzdesign eine ID und dessen englische und deutsche Form ausgelesen werden. Im Auszug sind nur Münzen enthalten, für die eine Beschreibung in beiden Sprachen vorliegt.

2.2 Die grundlegenden Arbeiten von Klinger, 2018 und Deligio und Gencer, 2021

In Klinger (2018) werden die englischen Daten der Münzbeschreibungen des *Corpus Nummorum Online* zu einer NLP-Analyse aufbereitet. Ziel von Klinger (2018) war die Implementierung von *Named Entity Recognition (NER)* und *Relation Extraction (RE)*. *NER* steht für das Erkennen von nominalen Bezeichnern für bestimmte Individuen (z.B. Namen wie *Athena*) in einem Text. Klinger (2018, S. 12) identifiziert nominale Bezeichner der vier Entitätengruppen PERSON, OBJEKT, ANIMAL und PLANT. PERSON enthält dabei alle Bezeichner von Menschen und Göttern, OBJEKT hingegen enthält Begriffe für unbelebte Gegenstände wie *bow*. ANIMAL enthält Bezeichner von tierischen Entitäten und PLANT enthält Bezeichner von pflanzlichen Entitäten.
RE steht für das Ermitteln der Beziehungen zwischen Entitäten. Diese können als Tripel der Form (NE_1, α, NE_2) modelliert werden, wobei NE_1 und NE_2 für Entitäten stehen und α für die Wörter zwischen den Entitäten (Bird, Klein und Loper (2009)). Am folgenden Beispiel aus Klinger (2018, S. 12–14) sieht man die Schritte der Analyse für diese beiden Aufgaben.

Apollo seated left on omphalos, holding bow in right hand.

Nach der NER Analyse wurden die Entitäten *Apollo*, *omphalos* und *bow* entdeckt. *Apollo* ist als PERSON identifiziert worden und *omphalos* und *bow* als OBJECT. *RE* wird von Klinger (2018) nur für (PERSON, OBJEKT) Paare ermittelt.

Daraus werden zwei (Person, Objekt) Paare erstellt: (Apollo, omphalos) und (Apollo, bow). Diese werden dann mit dem Originalsatz an die *RE*-Einheit weitergereicht, welche die Relation

zwischen den Entitäten ermittelt:

[(„Apollo“,„seated_on“,„omphalos“),(„Apollo“,„holding“,„bow“)]

Apollo und *omphalos* sind durch *seated_on* miteinander verbunden und *Apollo* ist durch *holding* mit *bow* verbunden.

Deligio und Gencer (2021) erweitert die Arbeit von Klinger (2018) um folgende Punkte: Es werden zusätzlich Paare von allen Entitätengruppen im *RE* extrahiert. Ebenfalls soll das Modell auf die deutschen Münzbeschreibungen angewendet werden. Hierfür wurden die Tabellen der Entitätsgruppen manuell übersetzt. Die Münzbeschreibungen liegen allerdings bereits in Deutsch vor (Deligio und Gencer (2021, S. 64)). Zusätzlich zu der *Relation Extraction* zwischen zwei Entitäten und einem Verb, sollen auch *Relations* mit nur einer Entität und einem Verb (NE_1, α) ermittelt werden, da viele Entität-Verb-Beziehungen kein Objekt besitzen.
Zur Ermittlung der Relationen werden die Verben in verschiedene Klassen unterteilt, welche eine semantische Kategorie der Verben definiert z.B. alle Verben die ausdrücken, dass etwas gehalten wird (Deligio und Gencer (2021, S. 38)). Die Entitätenklassen, die Klinger (2018) bereits erstellt hat, werden um eine neue Tabelle namens VERBS erweitert. Dies ist notwendig, da Verben oft nicht als solche erkannt werden, sondern als Nomen oder Adjektiv identifiziert werden.
Verschiedene Modelle und Funktionen der python-Library *SpaCy* werden als Grundlage für die Implementierung verwendet (Deligio und Gencer (2021, S. 24–25)). So wird *NER* durch das vortrainierte Modell *EntityRecognizer* implementiert, welches auf das Erkennen der Entitätsgruppen trainiert wird. Mithilfe des *DependencyParser* werden Abhängigkeiten der Wörter innerhalb eines Satzes erkannt. Sätze werden durch den *Tokenizer* in Einzelwörter und Satzzeichen zerteilt und einem Wort kann ein *Part of Speech Tag* zugeordnet werden, welcher die Wortart angibt.

Bei der Untersuchung der deutschen und der englischen Modelle fallen folgende Sachverhalte auf. Im Vergleich der deutschen und englischen Münzbeschreibungen hebt Deligio und Gencer (2021, S. 76–78) hervor, dass im Englischen oft Verben im Partizip Perfekt (*„Veiled [...] Cybele“*) verwendet werden, wobei im Deutschen diese Verbform komplett wegfällt (*„Kybele[...] mit Schleier“*). Die dadurch ausgedrückte Entität wird somit nur im Deutschen erkannt, da die Partizipform nicht erkannt wird. Deligio und Gencer (2021, S. 78) beschreiben für das Deutsche auch die Schwierigkeit der Analyse von Komposita wie *Heraklesknabe* und *Widderkopf*, welche nicht erkannt werden, obwohl *Herakles* und *Widder* als Entität verzeichnet ist. Die Annotationsfunktion ist nicht in der Lage zusammengesetzte Worte zu erkennen. Dies tritt im Englischen nicht auf, da dort Konstruktionen wie *Infant Heracles* bevorzugt werden, bei denen kein Verbinden mehrerer Worte auftritt.
Eine besondere Herausforderung stellt das deutsche Plural- und Kasussystem dar. In den Tabel-

len der Entitätengruppen steht nur eine Form eines Wortes, was bedeutet das sehr viele andere Formen nicht erkannt werden. Die Entitätentabellen wurden dementsprechend um die möglichen Formen der Worte manuell erweitert. Im Englischen ist dies aufgrund der geringeren Vielfalt des Plural- und Kasussystems ein kleineres Problem. In Abschnitt 2.3 werden die möglichen Flektionsformen der beiden Sprachen vorgestellt.

2.3 Flektion im Deutschen und Englischen

Flektion ist definiert als die Anpassung eines Wortes an bestimmte Merkmale einer Merkmalsklasse (vgl. Michel (2020, S. 43)) wie Ausdrücken des Plurals. Bei Verben nennt man dies Konjugation, und bei anderen Wortarten, wie Nomen und Adjektiven, Deklination.

2.3.1 Deutsche Flektion

Nomen werden im Deutschen nach Kasus und Numerus dekliniert. Hierfür wird für gewöhnlich ein Suffix an den Wortstamm gefügt und in manchen Fällen der Stammvokal geändert (sog. Ablaut). Das Genus (maskulin, feminin, neuter) eines Nomen ist für gewöhnlich nicht änderbar (vgl. Michel (2020, S. 47)). Im Kasus unterscheidet man zwischen 4 Fällen: Nominativ, Genitiv, Dativ und Akkusativ, wobei es drei verschiedene Deklinationsschemas gibt. Die Zuordnung eines Wortes zu einem dieser Deklinationsschema ist nicht an der äußeren Form des Wortes erkennbar und muss für jedes Wort erlernt werden (vgl. Michel (2020, S. 47)).
Das Deutsche besitzt zudem zwei Numeri, Singular und Plural. Zur Bildung des Plurals werden 9 verschiedene Strategien beschrieben (vgl. Michel (2020, S. 13)):

Tabelle 1: Strategien zur Bildung des Plurals im Deutschen

	Nom. Singular	**Nom. Plural**
ø	der Adler	die Adler
-s	das Auto	die Autos
-er	das Schwert	die Schwerter
-e	der Speer	die Speere
Umlaut	der Garten	die Gärten
-en	der Automat	die Automaten
-n	die Münze	die Münzen
Umlaut + e	der Plan	die Pläne
Umlaut + er	das Haus	die Häuser

Hinzu kommen Sonderregeln für Fremdworte im Deutschen z.B. *Lexikon*(Sg.) und *Lexika*(Pl.). Wie beim Kasus ist keine regelmäßige Zuordnung der verschiedenen Strategien zu einem Wort möglich, sodass dies ebenfalls für jedes Wort erlernt werden muss.
Zudem existieren Fälle, in denen man nicht an der äußeren Form eines Wortes sehen kann, welchen Numerus und Kasus dieses Wort hat z.B. *Lehrer* und man stattdessen diese Informationen aus dem Kontext ableiten muss. In dem Satz *„Der Lehrer betritt das Klassenzimmer“* ist *Lehrer* Nominativ Singular, wohingegen in dem Satz *„Die Autos der Lehrer stehen auf dem Parkplatz“* dasselbe Wort im Genitiv Plural steht. Adjektive stimmen in Kasus, Numerus und Genus mit ihrem zugeordneten Nomen überein. Auch hier unterscheidet man zwischen drei Deklinationsschemas, welche ebenfalls unregelmäßig verteilt sind.

Durch diese vielen verschiedenen, unregelmäßigen Strategien der Deklination fällt es schwer eine flektierte Form auf eine Grundform zurückzuführen. Ein Algorithmus muss bereits Informationen zu einem konkreten Wort besitzen, und kann nicht zuverlässig aus gelernten Regeln herleiten, zu welchem Deklinationsschema ein Wort gehört. Die Folge ist, dass die Anzahl der Wörter, über die diese Informationen angegeben wird, Auswirkungen auf die Qualität der Ergebnisse haben wird.

Die Flektion von Verben wird als Konjugation bezeichnet. Verben werden nach Numerus (Singular/Plural), Tempus, Modus, Person und Genus Verbi (Aktiv/Passiv) konjugiert (vgl. Michel (2020, S. 44–46)). Im Zuge dieser Arbeit werden nicht alle möglichen Verbformen betrachtet werden. Die Beschreibungen der Münzdaten sind in 3. Ps Indikativ Präsenz verfasst, wobei Genus Verbi und Numerus variabel sind. Verben werden in zwei Konjugationsklassen geteilt. Bei der Klasse der schwachen Verben werden die Formen regelmäßig durch Suffixe und Zirkumfixe ausgedrückt. Das schwache Verb *spielen* hat im Präteritum die Form *spiel* + *t* und als Partizip II *ge* + *spiel* + *t*. Bei starken Verben werden die Formen durch Änderung des Stammvokals ausgedrückt. Wie der Stammvokal verändert wird, folgt allerdings keiner bestimmten Regel, sondern muss für die entsprechenden Verben gelernt werden. Das starke Verb *singen* ändert den Stammvokal: *sang* und *ge* + *sung* + *en*. Partizipien stellen einen Sonderfall dar, da diese auch als Adjektiv verwendet werden können und somit eine besondere Schwierigkeit in der korrekten Identifizierung der Wörter darstellen (vgl. Imo (2016, S. 44)). Formen mit *-end*, wie *haltend*, drücken Partizip I aus, welcher immer als Adjektiv verwendet wird (vgl. Imo (2016, S. 55)). Partizip II wird durch Präfix *ge-* und den Suffixen *-en* oder *-t* ausgedrückt, wie *gestützt*, und kann sowohl als Adjektiv, als auch als Teil eines Verbgefüges (vgl. Imo (2016, S. 55)) verwendet werden. Diese Partizipform findet sich oft in den Münzbeschreibungden des *Corpus Nummorum Online* und muss entsprechend im Satzkontext die korrekte Wortart zugewiesen bekommen.

Abgesehen von Flektion, in der ein Wort mit einem Affix verbunden wird um eine grammatische Funktion auszudrücken, können Affixe und Wortstämme miteinander kombiniert werden, um

Wortart doer Bedeutung eins Wortes zu ändern. Dies nennt man Derivation. Die Verbindung zwei oder mehrere Wortstämme wird als Komposition bezeichnet und tritt auch in den Daten des des *Corpus Nummorum Online* auf. In zwei Drittel aller Komposita des Deutschen sind die Stämme direkt miteinander verbunden, in den übrigen Fällen wird ein sogenanntes Fugenelement verwendet z.B. „Arbeit“ +„s“ + „Amt“ ergibt „Arbeitsamt“. Bei Komposita gilt generell das Kopf-Rechts-Prinzip (vgl. Michel (2020, S. 15)). Das bedeutet, dass das am weitesten rechts stehende Glied des Kompositums den Kasus, Numerus, Genus und Bedeutungskategorie des Gesamtwortes bestimmt. So beschreibt das Wort „Löwenfell“ eine Art von Fell und keine Art von Löwe. Diese Information ist in der Analyse von Komposita im Corpus Nummorum äußerst wichtig, falls die Bestandteile eines Kompositums zu verschiedenen Entitätenkategorien gehören. Einem Wort kann nur eine Entitätenkategorie zugeordnet sein, und nach Kopf-Rechts-Prinzip ist die Entität des am weitesten rechts stehenden Elements entscheidend.

2.3.2 Englische Flektion

Im Englischen stellen sich ähnliche Schwierigkeiten wie im Deutschen. Auch hier wird Deklination und Konjugation durch das Anhängen von Suffixen am Wortstamm ausgedrückt. Im Gegensatz zum Deutschen, wird im Englischen nur nach Kasus und Numerus dekliniert. Es existiert kein Genus. Auch wird im Englischen nur der Genitiv markiert (vgl. Huddleston und Pullum (2002, S. 1595 – 1596)). Dies erfolgt in der Regel durch Affigieren von „’s “ an den Wortstamm z.B. „man“ wird zu „man’s“. Nur wenn das Wort im Plural auf „s“ endet, wird Genitiv Plural durch reines Anhängen des Apostrophs ’ ausgedrückt z.B. „dogs“ wird zu “dogs’ “. Zur Markierung des Numerus unterscheidet man zwischen regelmäßigen und unregelmäßigen Wörtern (vgl. Huddleston und Pullum (2002, S. 1585–1586)). Die regelmäßige Bildung des Plurals wird durch Anhängen des Suffixs „-s“ an den Wortstamm gebildet, wobei einige Sonderfälle auftreten, welche allerdings oft vorhersagbar sind, wie Tabelle 2 zeigt:

Tabelle 2: Orthografische Variationen des -s Plurals im Englischen

-s	**-ies**	**-es**	**Doppelung des letzten Konsonanten + es**
dog - dogs	cherry - cherries	echo - echoes	quiz - quizzes

Der Plural kann auch irregulär durch Änderung des letzten Konsonanten des Stammes markiert werden z.B. „knife“ und „knives“. Diese Änderung ist allerdings nicht vorhersagbar und trifft nicht auf alle Worte mit diesen Endkonsonanten zu (vgl. Huddleston und Pullum (2002, S. 1587–1590)). Auch wenn im Englischen wie im Deutschen zahlreiche unregelmäßige Formen existieren, die nicht am Aussehen des Wortes erkennbar sind, besitzt Englisch weniger Strategien zur Deklination sowie nur einen markierten Kasus. Dies erleichtert das Erkennen einer Grundform anhand einer flektierten Form enorm.

Konjugation im Englischen hat ebenfalls weniger mögliche Formen als im Deutschen. Verben

haben mindestens 3, und maximal 5 verschiedene konjugierte Formen, die in Abbildung 1 gezeigt werden.

Abbildung 1: Konjugierte Verbformen in Englisch (vgl. Huddleston und Pullum (2002, S. 1596))

PLAIN FORM	PLAIN PRESENT	PRETERITE	PAST PARTICIPLE	3RD SG PRESENT	GERUND-PARTICIPLE
take		*took*	*taken*	*takes*	*taking*
love		*loved*		*loves*	*loving*
cut				*cuts*	*cutting*

Partizipien in Englisch werden durch Anhängen von *-ing* oder *-ed* an den Wortstamm gebildet (vgl. Huddleston und Pullum (2002, S. 78–79)). Der zweite fall stellt das Partizip Perfekt dar und wird in den Münzbeschreibungen des *Corpus Nummorum Online* verwendet, um dadurch das Tragen oder Besitzen von Objekten in adjektivischer Form auszudrücken z.B. *„Helmeted“* bedeutet, dass ein Helm getragen wird. Partizipien mit *-ing* wie *„Sailing“* werden Gerund oder Partizip Präsenz genannt und können sowohl adjektivisch, als auch als Nominalisierung des Verbs verwendet werden (vgl. Huddleston und Pullum (2002, S. 80–81)). Welcher Fall vorliegt, muss durch den Satzkontext erschlossen werden.
In welche Klasse welches Wort gehört ist auch hier nicht regelmäßig, wobei dennoch einige Tendenzen möglich sind. Auch hier bietet sich, im Vergleich zur deutschen Konjugation, der Vorteil, dass weniger Suffixe zur Markierung der Person existieren.

Zusammenfassend lässt sich sagen, dass die Flektion sowohl im Deutschen und Englischen viele Ausprägungen hat und es nicht immer ersichtlich ist, auf welches Wort welche Strategie angewandt werden darf. Im Deutschen ist die Flektion allerdings wesentlich komplexer durch die höhere Anzahl an Kasusmarkierungen und Verbmarkierungen. Die Grundform eines Wortes anhand seiner flektieren Form zu erkennen, ist folglich nicht leicht. Zunächst muss sich auf eine Definition von Grundform geeinigt werden. Zwei Varianten davon sind das Lemma und der Stamm. Sie werden mithilfe der NLP-Methoden Lemmatisierung und Stemming ermittelt, die in Abschnitt 2.4 vorgestellt werden.

2.4 Einführung in Lemmatisierung und Stemming

Dieser Abschnitt beschäftigt sich mit Lemmatisierung und Stemming. Die beiden Begriffe werden zunächst definiert und die Unterschiede zwischen den beiden Prozessen hervorgehoben. Anschließend wird eine mögliche Implementierung von Lemmatisierung für Englisch und Deutsch besprochen. Danach werden Algorithmen für Stemming im Englischen, und Algorithmen für Stemming im Deutschen vorgestellt.

Ziel von Lemmatisierung und Stemming ist das Zurückführen eines flektierten Wortes auf eine

Grundform (Manning, Raghavan und Schütze (2009, S. 32)), beispielsweise das Zurückführen von *„organize“*, *„organizes“*und *„organization“*auf dieselbe Grundform. Dies ist sehr nützlich für *information extraction*. Der User gibt eine Anfrage ein und die vorhandenen Texte werden nach den Wörtern dieser Query durchsucht. Die genaue Form der Wörter soll hierbei keine Rolle spielen (Manning, Raghavan und Schütze (2009, S. 32)).

2.4.1 Lemmatisierung

Nach Manning, Raghavan und Schütze (2009, S. 32) ist Lemmatisierung die angemessenere Variante der Reduktion eines Wortes auf eine Grundform (*„doing things properly“* (Manning, Raghavan und Schütze (2009, S. 32))). Hierbei wird die Verwendung eines Wortes im Satz analysiert und eine morphologische Analyse durchgeführt, sodass lediglich die Flektionssuffixe entfernt werden. Als Grundform bleibt dann die lexikalische Form des Wortes übrig. Die Grundform ist hierbei immer ein real existierendes Wort. Im Beispiel von *„organize“*, *„organizes“*und *„organization“* ,wird *„organize“* als Grundform angenommen und *„organizes“* durch Entfernung des Flektionssuffixes -s auf diese zurückgeführt. *„organization“* hingegen kann nicht auf *„organize“* zurückgeführt werden, da es sich bei -ation um ein derivationelles Suffix handelt. Implementiert wird Lemmatisierung häufig nur durch ein Nachschlagen des Wortes in einem Wörterbuch.

Die bereits in den vorherigen Arbeiten (Klinger (2018), Deligio und Gencer (2021)) verwendete Bibliothek SpaCy für Python besitzt bereits Lemmatisierungsfunktionen für u.a. Englisch und Deutsch.
Für das Englische wird ein sog. *Rule-Based Approach* [4] dafür angewendet. Dabei werden Informationen zur Wortart, ein Wörterbuch (sog. Index), Ersetzungsregeln von Suffixen und eine Liste von Spezialfällen verwendet [5]. Für ein Wort wird zuerst die Wortart (Part of Speech Tag) ermittelt, da die Ersetzungsregeln nach Wortarten sortiert sind. Dann wird ermittelt, welches Suffix vorliegt, und dieses wird mit der dazugehörigen Angabe ersetzt. Wenn das dadurch entstandene Wort im internen Wörterbuch vorhanden ist, haben wir die Grundform gefunden. Ansonsten wird geprüft, ob das Wort zu den Spezialfällen zählt, für die besondere Regeln gelten.
Für das Deutsche wird ein sog. *Lookup-Based Approach* angewendet. Hierbei wird lediglich eine Liste von Wörtern und deren verschiedene Wortformen verwendet. Für ein Wort, das lemmatisiert werden soll, wird dann diese Liste nach dieser Wortform durchsucht, und die dazugehörige Grundform ausgegeben. Es wird kein Gebrauch von der Wortart gemacht.
Die Library *NLTK* bietet ebenfalls Lemmatisierer für Englisch an. Ein Lemmatisierer von *NLTK* heißt *Wordnet*[6] und verwendet ebenfalls einen *Rule-Based Approach*. Als Teil des Hannover-Taggers (HanTa) wird für das Deutsche ein *rule-based* Lemmatisierer angeboten (Wartena

[4] https://spacy.io/api/lemmatizer (01.02.2022)
[5] https://github.com/explosion/spacy-lookups-data/tree/master/spacy_lookups_data/data (01.02.2022)
[6] https://www.nltk.org/_modules/nltk/stem/wordnet.html (20.03.2022)

(2019)). Diese beiden Lemmatisierer werden in Abschnitt *Abschnitt 4* mit denen von SpaCy verglichen.

2.4.2 Stemming

Stemming ist ein heuristischer Prozess, der Wortenden abtrennt und somit auch derivationelle Affixe entfernen kann (Manning, Raghavan und Schütze (2009, S. 32)). Das Resultat muss, im Gegensatz zu Lemmatisierung, kein real existierendes Wort sein. Die Grundform kann nach Manning, Raghavan und Schütze (2009, S. 33) als eine Äquivalenzklasse verstanden werden. Im Beispiel von *„organize“*, *„organizes“* und *„organization“* wird *„organ“* als Grundform aller drei Wörter ermittelt, da -ize, -izes und -ization als kombinierte Suffixe erkannt werden. Algorithmen für Stemming bestehen aus sprach-spezifischen Regeln, verwenden aber keine Informationen über die Funktion des Wortes im Satz (z.B. Part-of-Speech Tag).
Der bekannteste Stemming Algorithmus für das Englische ist Porters Algorithmus (Porter (1980)). Dieser Algorithmus wurde nach seiner Veröffentlichung noch weiter entwickelt und wird als Porter2 oder auch als *Snowball*[7] bezeichnet. Dieser löscht in mehreren Phasen sukzessiv die Suffixe der Wörter. In jeder Phase wird nur nach bestimmten Suffixen gesucht, wobei zuerst flektionale und danach derivationelle Suffixe berücksichtigt werden. In jeder Phase gibt es zudem bestimmte Regeln, die eingehalten werden müssen. Diese Regeln sind oft kontextabhängig und werden auf bestimmte Bereiche des Wortes angewandt wie die erste Vokal-Konsonanten Abfolge im Wort. Einige der Regeln besagen auch, dass immer das längste gefundene Suffix gelöscht werden soll (Manning, Raghavan und Schütze (2009, S. 33)), oder wenn das Wort nur noch eine Silbe lang ist, es nicht weiter verkürzt werden darf.
SpaCy bietet eine Implementierung von *Snowball* für Englisch an.
Ein alternativer Stemmingalgorithmus, ist der Stemmer von Krovetz. Dieser Stemmer wendet Wissen über Flektion und Derivation des Englischen an (vgl. Krovetz (1993)) um so Grundformen zu ermitteln.

Auch für das Deutsche existiert ein *Snowball*-Stemmer, welcher als *Pystem*[8] in Python verfügbar ist. Allerdings übertrifft der in Weissweiler und Fraser (2018) vorgestellte Stemmingalgorithmus *Cistem* diesen in den dort geprüften Aufgaben, weshalb ihm hier Vorrang gegeben wird.
In Weissweiler und Fraser (2018) werden fünf verschiedene Stemmer für das Deutsche zunächst getestet und der mit den besten Ergebnissen ausgewählt. Als bester Stemmingalgorithmus hat sich der Algorithmus von Caumanns (1997) erwiesen. Dieser Algorithmus entfernt rekursiv die Zeichenfolgen *e, s, n, t, em, er* und *nd*, da diese Zeichen als Bestandteile aller Flektionssuffixe identifiziert worden sind (Caumanns (1997, S. 4–6)). Zudem werden einige Zeichen und Zeichenabfolgen eines Wortes mit Sonderzeichen substituiert, um das fälschliche Löschen dieser zu verhindern. Weissweiler und Fraser (2018, S. 90, 91) verbessern diesen Algorithmus durch

[7] https://snowballstem.org/algorithms/english/stemmer.html (03.02.2022)
[8] https://github.com/snowballstem/pystemmer (20.03.2022)

das Einführen einer festen Reihenfolge der Schritte und Anpassung der Substitutionsregeln. Der Algorithmus von Weissweiler und Fraser (2018) wird von *NLTK* zur Verfügung gestellt.
In Abschnitt *Abschnitt 4* werden die Ergebnisse der Stemmer *Snowball* und *Cistem* mit den Ergebnissen des *korvetz-Stemmer* und *Pystem* verglichen.

2.5 Verwendete Metriken

Im Bezug auf die Annotation werden ein reiner Lemmatisierungsansatz, ein reiner Stemmingansatz, ein hybrider Lemmatisierung- und Stemmingansatz und die Annotation von Deligio und Gencer (2021) miteinander verglichen. Für den Vergleich zweier Modell werden dabei folgende Metriken verwendet, welche im Kontext von *Machine Learning* zur Bewertung genutzt werden (vgl. Bird, Klein und Loper (2009, Chapter 6.3)). Üblicherweise wird dort ein neues Modell durch den Vergleich mit einem Goldstandard bewertet. Ein Goldstandard ist ein Modell von dem bewiesen ist, dass immer korrekte Ergebnisse bereitstellt. Hierbei werden vier verschiedene Werte betrachtet. Bei Klassifikationsmodellen stehen **True** und **False** für die Bewertung eines Elements von dem Goldstandard d.h. richtige und falsche Vorhersagen des Modells. Die Kategorien **predicted True** und **predicted False** hingegen stehen für die richtigen und falschen Vorhersagen des Modells, das bewertet wird (Bird, Klein und Loper (2009, Chapter 6.3)).

Tabelle 3: Übersicht Metriken nach Manning, Raghavan und Schütze (2009, S. 155)

	True	**False**
predicted True	True Positive (TP)	False Positive (FP)
predicted False	False Negative (FN)	True Negative (TN)

An die Stelle des Begriffs der Vorhersage tritt hier die gewonnene Annotation im Vergleich zur Annotation von Deligio und Gencer (2021), welche die Rolle des Goldstandards übernimmt. Dabei ist zu beachten, dass für diese Annotation weder Vollständigkeit noch Fehlerfreiheit bewiesen ist. Deshalb werden die Ergebnisse zusätzlich manuell überprüft.
Die Kategorie **True** bedeutet deshalb hier, dass ein Wort mit einer bestimmten Entitätenkategorie annotiert wurde und **False** bedeutet, dass es nicht mit dieser Entitätenkategorie annotiert wurde. Unter *True Positive* ist somit eine Übereinstimmung der gewonnenen Annotation mit dem Goldstandard zu verstehen; unter *False Positive* sind Fälle einzuordnen, die in der gewonnenen Annotation auftreten, im Goldstandard aber nicht vorhanden sind. Die Erhebung der Anzahl an Fällen pro Kategorie ermöglicht es uns, die zwei Kennwerte *Precision* und *Recall* zu berechnen, die wir hier als Metrik heranziehen möchten. *Recall* beschreibt hierbei, wie viel Prozent der annotierten Elemente des Goldstandards gefunden wurden. : $\frac{TP}{TP+FN}$ (Bird, Klein und Loper (2009, Chapter 6.3)). Ein hoher *Recall* bedeutet also, dass sehr viele annotierte Wörter auch von dem Goldstandard annotiert wurden.
Precision steht hier dafür wie viel Prozent der neu annotierten Elemente auch von dem Goldstandard annotiert wurden : $\frac{TP}{TP+FP}$ (Bird, Klein und Loper (2009, Chapter 6.3)). Eine hohe *Precision* besagt also, dass nur wenige weitere Wörter annotiert wurden, die nicht auch vom

Goldstandard annotiert wurden.
Da kein objektiver Goldstandard für die verschiedenen Strategien zur Annotation existiert, werden diese Werte genutzt, um die Modelle miteinander zu vergleichen. Aus diesem Grund muss nach der Berechnung von *Recall* und *Precision* manuell geprüft werden, welche Entitäten nur von einem, aber nicht von dem anderen Modell annotiert wurden. Durch diese manuelle Prüfung kann bewertet werden, welcher der Algorithmen, die miteinander verglichen werden, die meisten korrekte Annotationen erzeugt.

3 Modellierung

Im folgenden Kapitel werden die Schwierigkeiten einer Annotation mittels Lemmatisierung und Stemming beschrieben und mögliche Lösungen vorgestellt. Darauf aufbauend wird gezeigt, wie der Ablauf des neuen Annotierungsalgorithmus aussieht und wie zuvor erdachten Lösungsansätze implementiert werden.

3.1 Problembeschreibung

In der Arbeit von Deligio und Gencer (2021) wurden die verschiedenen flektierten Formen der Wörter manuell als *alternative names* in den SQL-Datendump des *Corpus Nummorum Online* eingetragen. Folgende Problembereiche haben sich dabei herauskristallisiert. Es erfordert zusätzliche Arbeit für jede Wortform alle flektierten Formen zu ermitteln und aufzuschreiben. Ebenfalls sind die Einträge anfällig für Tippfehler, da die Wörter manuell eingetragen werden. So wurde beispielsweise „Bienestöcke“ anstatt „Bienenstöcke“ eingetragen. Zudem ist nicht garantiert, dass man an jede mögliche Form gedacht hat, die ein Wort haben kann. So fehlt zur Grundform “Baum“ die Dativ- und Akkusativ-Plural-Form „Bäumen“, welche allerdings in den Münzbeschreibungen vorkommt *(„Tetrastyler Rundtempel zwischen zwei Bäumen; darin Apollon (Iatros) mit Schlangenstab und kleiner Eros mit erhobener Rechten.“(ID: 3377))*. Dadurch sind die Einträge unvollständig.

Durch ein automatisiertes Erkennen der Grundform durch eine NLP-Analyse (Lemmatisierung & Stemming) eines flektierten Wortes können diese Probleme gelöst werden. Hierfür müssen die NLP-analysierten Wörter der Beschreibungen der Münzen mit den NLP-analysierten Einträgen der Datenbank verglichen werden. Wenn eine Übereinstimmung gefunden wird, heißt dies, dass das Wort im Satz eine Entität ist, und das das Wort mit dem dazugehörigen Entitätenlabel wird gespeichert. Der Aufbau der Annotationsfunktion soll hierbei der Funktion aus Deligio und Gencer (2021) folgen. Der Input besteht aus den Beschreibungen der Münzdaten (sog. *Designs*) und den 5 verschiedenen Entitätenlisten (PERSON, ANIMAL, OBJECT, PLANT, VERB) für Deutsch und Englisch. Der Output der Annotationsfunktion besteht aus einer Liste von Tupeln der Art (vordere Wortgrenze als Position im Satz, hintere Wortgrenze als Position im Satz, Entitätenlabel). Das Wort *Löwe* kann beispielsweise die Annotation (0,5,ANIMAL) erhalten.

Angenommen wir betrachten die Entitätsklasse der Verben, welche die Wörter *stehen* und *halten* enthält, und wollen das folgende *Design* annotieren:

„*Nackter Dionysos stehend von vorn, Kopf nach links, in der vorgestreckten Rechten Weintraube, im linken Arm Gewand und zwei Speere haltend. Kurze Bildleiste.“(ID: 399)*

Tabelle 4: Vergleich von Lemmatisierung und Stemming auf zwei Verben

	Ergebnis Lemmatisierung	**Ergebnis Stemming**
halten	halten	hal
haltend	halten	hal
stehen	stehen	steh
stehend	stehen	steh

Tabelle 4 zeigt die lemmatisierten und gestemmten Formen der Verben *stehend* und *haltend* in Partizip I und Infinitiv Präsenz. Man sieht, dass die beiden Fomren auf dieselben Lemmas bzw. Stems zurückgeführt werden. *stehend* und *haltend* werden somit erfolgreich als Formen der beiden Einträge der Entitätsdatenbank erkannt und als VERB annotiert. Wie man am Ergebnis des Stemmings sieht, werden die Wörter der Münzbeschreibungen nicht zwangsläufig auf die Form in der Datenbank zurückgeführt. Wichtig ist nur, dass sowohl die Entität als auch das Wort aus dem Satz auf dieselbe Grundform zurückgeführt werden.

Wie bereits in Deligio und Gencer (2021, S. 78) besprochen, finden sich in den deutschen Beschreibungen einige Komposita, welche aus einer oder mehreren zusammengefügten Entitäten bestehen, wie *Heraklesknabe, Widderkopf, Löwenfell, Elefantenzahn* und *Dionysoskind*. Um diese Worte korrekt zu analysieren, müssen sie zunächst in ihre Bestandteile geteilt werden. Dann muss jedes Teilwort NLP-analysiert werden und mit den NLP-analysierten Entitäten verglichen werden. In Fällen wie *Heraklesknabe* wird nur ein Teilwort erkannt, und die Entität des Teilwortes kann auf das gesamte Wort übertragen werden. Dies ist allerdings nicht immer so einfach. Der Begriff *Löwenfell* enthält beispielsweise zwei verschiedene Entitäten. das Teilwort *Löwe* ist vom Typ ANIMAL, wohingegen *Fell* vom Typ OBJECT ist. Nach der Kopf-Rechts-Regel des Deutschen bestimmt das am weitesten rechts stehende Wort den lexikalischen Typ des Gesamtwortes. Folglich müssen die gefundenen Entitäten der Teilworte eines Kompositums miteinander verglichen, und das am weitesten rechts Stehende ermittelt werden. Dessen Entität wird dann auf das gesamte Kompositum übertragen. Das Wort *Löwenfell* erhält also die Annotation OBJECT.

Auch Satzzeichen in Entitäteneinträgen wie das Zeichen - in *Asklepios-Schlange* bilden ein Hindernis für Lemma-tisierung- und Stemmingalgorithmen. Diese analysieren ein Wort nur bis zum

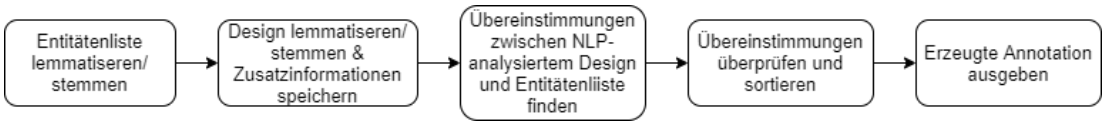
Satzzeichen und schneiden den Rest einfach ab. Deswegen müssen vor der Analyse Satzzeichen entfernt werden. Da allerdings die Koordianten des Wortes im unveränderten Originalsatz benötigt werden, muss diese Information gespeichert werden.
Einige Entitäten finden sich außerdem in mehreren Entitätenlisten. Die Entitäten *Bienenstock* und *Ast* sind sowohl als OBJECT als auch als PLANT eingetragen, und *Beehive* ist zweifach in OBJECT und ANIMAL eingetragen. In diesen Fällen muss definiert werden, welcher der beiden Label Präzedenz hat.

Ein weiteres Problem entsteht durch die Rückführung verschiedener Wörter auf denselben Stamm oder dasselbe Lemma. So kann *Ruderer* auf *Ruder* zurückgeführt werden, oder *Hermes* wird auf *her* zurückgeführt. Im ersten Fall wird *Ruderer* dann als Entität vom Typ OBJEKT erkannt und im zweiten Fall wird das Pronomen *her* als PERSON getaggt. Beide Fälle sind falsche Annotationen und können durch einen verbesserten Lemmatisierer oder Stemmer vermieden werden. Ebenso kann die Wortart berücksichtigt werden, sodass beispielsweise Personalpronomen nicht annotiert werden.

3.2 Programmbeschreibung

Im Folgenden wird der Ablauf des Programms mit NLP-Methoden und die Ermittlung der Annotationen vorgestellt.

Abbildung 2: Ablauf des Programmes



Zu Beginn des Annotationsprozesses werden zuerst die Einträge der Entitätenliste mit der gewählten NLP-Methode verarbeitet. Dann werden die *Designs* nacheinander mit derselben Methode verarbeitet und zusätzliche Informationen zur Wortart und dem Ursprungswort gespeichert. Anschließend werden die Einträge der Eintitätenliste in dem *Design* gesucht und Übereinstimmungen zwischengespeichert. Nachdem alle möglichen Annotationen eines Satzes gesammelt sind, werden diese miteinander verglichen und Teile davon gegebenenfalls aussortiert. Zum Schluss wird die fertige Annotation in derselben Form, wie Deligio und Gencer (2021) ausgegeben.

Der Ablauf der Annotation der deutschen und der englischen Sätze unterscheidet sich dabei nicht voneinander. Die Annotationsfunktion bietet drei verschiedene Möglichkeiten der NLP-Analyse: Entweder werden die Wörter lemmatisiert, gestemmt oder beides. Im letzten Fall werden die Annotationsergebnisse der Lemmatisierung und des Stemmings separat erstellt und

miteinander vermischt. Beim Aufruf der Annotationsfunktion wird festgelegt, welche NLP-Methode verwendet wird.
Vor Beginn der Annotation müssen die Entitätslisten in der Datenbank, wie sie von Deligio und Gencer (2021) übernommen wurden, angepasst werden. Diese dürfen nicht mehrere Formen eines Wortes enthalten, sondern nur eine einzige. Hierfür werden die Einträge der Datenbank manuell angepasst. Tippfehler werden ebenfalls verbessert. Für Wörter, die in mehrere Entitätenklassen gehören können, wurde in der Datenbank eine weiter Spalte namens *multi_inheritance* erstellt. In dieser werden weitere Entitätentypen, die auf das entsprechende Wort zutreffen, festgehalten. Im Rahmen dieser Arbeit wird aber jedem Wort nur eine Entität zugeordnet und die Einträge von *multi_inheritance* nicht miteinbezogen.

Als erster Schritt des Annotationsprozesse werden die Entitätenlisten aus der Datenbank nacheinander eingelesen und für jeden Eintrag einer Liste das dazugehörige Ergebnis der gewählten NLP-Methode abgespeichert. Die neue Entitätenliste enthält die Originalformen der Einträge und die analysierten Formen. Falls das Originalwort in einem *Design* unverändert vorkommt, wird es so definitiv erkannt und ist nicht anfällig für Fehler des Algorithmus.
Die *Designs* werden einzeln nacheinander verarbeitet. Die Verarbeitung besteht daraus, dass alle Wörter des aktuellen Designs gestemmt oder lemmatisiert werden - je nach gewählter Methode - und das Ergebnis zu einer Zeichenkette zusammengefasst wird, dessen Reihenfolge mit der des ursprünglichen Satzes identisch ist. Zur Veranschaulichung wird der folgende Satz verarbeitet.

„Kaiser (Caracalla) auf Pferd nach rechts, vor ihm Tyche mit Asklepiosstatuette im rechten Arm.“(ID:5645)

Als NLP-Methode wird Lemmatisierung gewählt. Nach der Lemmatisierung der Wörter hat der Satz folgende Gestalt:
kaiser caracalla auf pferd nach rechts vor ich tyche mit asklepios statuette asklepiosstatuette im recht arm .

Auffällig ist, wie das Kompositum aufgelöst wurde. Die Teilworte *asklepios* und *statuette* wurden erfolgreich erkannt und getrennt in den Satz eingefügt. Zusätzlich wurde aber ebenfalls das Kompositum selbst eingefügt. Falls das Kompositum bereits als Entität in den Entitätenlisten eingetragen ist, kann es so direkt identifiziert werden.

Zusätzlich dazu wird das Ursprungswort des analysierten Wortes und dessen Part-of-Speech Tag, der mithilfe einer Funktion des Moduls SpaCy ermittelt wird, gespeichert. Dies ist notwendig, da nicht nur die Existenz einer Entität in einem Satz bestätigt werden soll, sondern auch dessen Satzkoordinaten. Mithilfe der gespeicherten Ursprungsform kann bei einer Übereisntimmung von einer Entität auf einem NLP-analysierten Wort des Satzes direkt die Koordinaten

ermittelt werden. Der Part-of-Speech Tag wird zwischengespeichert, da diese Information hilfreich ist, um einfache Fehler zu beseitigen, wie die fälschliche Annotation eines Nomens als Entitätentyp VERB. Komposita wie *Löwenfell* werden in diesem Schritt in ihre Bestandteile aufgeteilt. Dafür wird die Library *compound_split* eingesetzt. Diese berechnet die Wahrscheinlichkeit für Punkte, an denen ein Wort in Teilworte aufgespalten werden. In dieser Arbeit werden nur Wahrscheinlichkeiten über 98 % berücksichtigt. Anschließend wird jedes Teilwort lemmatisiert oder gestemmt und als Einzelwort in den Ausgabestring eingefügt. Dem analysierten Teilwort wird allerdings das Gesamtwort als Ursprungswort und eine Prüfzahl zugeordnet, wobei die kleinste Zahl das am weitesten links stehenste, und die größte Zahl das am weitesten rechts stehende Teilwort markiert. Damit kann später, wenn alle Wörter annotiert wurden, schnell ermittelt werden, welche Annotation dem gesamten Kompositum gegeben werden soll.
Für den Beispielsatz sieht die Zuordnung der lemmatisierten Formen zu ihrer Grundform folgendermaßen aus:

kaiser: [[NOUN, Kaiser]], " : [[PUNCT, (], [PUNCT,)], [PUNCT, ,]], caracalla: [[PROPN, Caracalla]], auf: [[ADP, auf]], pferd: [[NOUN, Pferd]], nach: [[ADP, nach]], rechts: [[ADV, rechts]], vor: [[ADP, vor]], ich: [[PRON, ihm]], tyche: [[NOUN, Tyche]], mit: [[ADP, mit]], asklepios: [[NOUN, Asklepiosstatuette, 0]], statuette: [[NOUN, Asklepiosstatuette, 1]], asklepiosstatuette: [[NOUN, Asklepiosstatuette, 2]], im: [[ADP, im]], recht: [[ADJ, rechten]], arm: [[NOUN, Arm]], .: [[PUNCT, .]]

Jedem lemmatisiertem Wort sind alle dazugehörigen Originalworte und Part-of-Speech Tags zugewiesen. Man beachte bei den Formen des Kompositums die ansteigende Prüfzahl.

asklepios: [[NOUN, Asklepiosstatuette, 0]],statuette: [[NOUN, Asklepiosstatuette, 1]], asklepiosstatuette: [[NOUN, Asklepiosstatuette, 2]].

Das Gesamtwort besitzt die höchste Wertigkeit und wird somit als wichtiger gewertet als die Teilworte.
Im nächsten Schritt werden die erweiterten Entitätenlisten und der NLP-analysierte Satz miteinander verglichen. Dazu wird ein regulrärer Ausdruckv erwendet, der alle Entitäten eienr Entitätenkategoire enthält, und diese im *Design* sucht. Das Ergebnis dieser Suche besteht aus allen Entitäten, die in diesem Satz gefunden wurden. Hierbei können überlappende Entitäten gefunden werden z.B. können für die Phrase *palm branches* sowohl nur *branch* als auch die gesamte Phrase als Entität gefunden werden. In einem späteren Schritt werden diese Überlappungen entfernt.

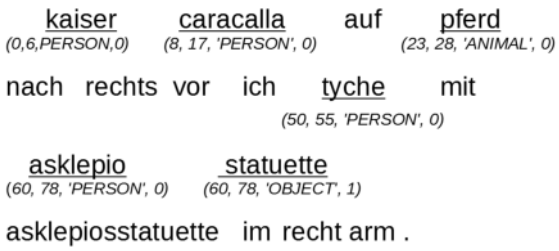
Nachdem die Entitäten einer Entitätsklasse in einem Satz ermittelt wurden, müssen die Satzkoordinaten der NLP-analysierten Wörter im unveränderten *Design* gefunden werden. Für Einzelwörter kann direkt das vorher abgespeicherte Ursprungswort abgelesen werden, und dessen

Koordinate im Satz direkt gefunden werden. Für gefundene Phrasen wie *palm branches* wird für jedes Einzelwort die Koordinaten ermittelt, und dann die größte und kleinste Koordinate als Ergebnis ausgegeben. Falls ein Teil der Phrase mehrfach im Satz vorkommt, muss darauf geachtet werden, dass die Koordianten des Phrasenteils innnerhalb der Phrase gewählt werden. Bei Komposita wird für ein gefundes Teilwort das gesamte ursprüngliche Kompositum als Ursprungswort zugeordnet, und dessen Koordinaten ausgewählt. Zudem wird die Prüfzahl gespeichert, welche angibt, wie weit rechts das Wort im Kompositum steht. Falls zu einem späteren Zeitpunkt ein anderes Teilwort als Entität identifiziert wird, kann mithilfe dieses Wertes geprüft werden, ob ein weiter rechts stehendes Teilwort gefunden wurde. wenn dieser Fall eintritt, wird das Entitätenlabel und die Prüfzahl des Gesamtwortes angepasst.
Die Wörter, die nicht Teil eines Kompositums sind, erhalten hier die Prüfzahl 0, damit alle Annotationen dasselbe Schema haben. Die Prüfzahlen werden nach fertigstellung der Annotation wieder entfernt, da sie dann nicht mehr benötigt werden.
Nachdem alle Entitätslisten geprüft wurden, hat die Annotation des Beispielsatzes folgende Gestalt:

[(50, 55, PERSON, 0), (0, 6, PERSON, 0), (60, 78, PERSON, 0), (8, 17, PERSON, 0), (60, 78, OBJECT, 1), (23, 28, ANIMAL, 0)].

Die ersten beiden Ziffern stehen für den Startpunkt und Endpunkt des Wortes im Originalsatz. Der dritte Eintrag beschreibt den Typ der gefundenen Entität und die letze Ziffer ist die Prüfzahl. Abbildung 3 zeigt, welches lemmatiserte Wort als eine Entität identifiziert wurde:

Abbildung 3: Annotation des Beispielsatzes



Dieses Beispiel zeigt, wie die Analyse eines Kompositums durchgeführt wird. Sowohl *asklepios* als auch *statuette* sind als Entitäten bekannt und werden korrekt identifiziert. Aus der bereits beschriebenen Zuordnung zum Originalwort *asklepios: [[NOUN, Asklepiosstatuette, 0]], statuette: [[NOUN, Asklepiosstatuette, 1]], asklepiosstatuette:[[NOUN, Asklepiosstatuette, 2]]* wird abgelesen, um welches Originalwort es sich handelt und welche Prüfzahl dem Teilwort zugeordnet ist. Anhand des Originalwortes werden dann die Startpunkte und Endpunkte ermittelt. Zu einem späteren Zeitpunkt wird dann bestimmt, welcher der beiden Einträge für *(*60,78)

vorrangig ist.

Im nächsten Schritt der Analyse werden die ermittelten Entitäten geprüft und Fehler bei diesen entfernt.
Die Entitätenliste zu dem Entitätstyp VERBS enthält ausschließlich Verben. Deshalb sollten auch nur Verben damit annotiert werden, und nicht fälschlicherweise andere Wortarten. Durch den abgespeicherten Part-of-Speech Tag zu jedem Wort eines Satzes kann man dies leicht prüfen. Zusätzlich zu Worten des Typs Verb, können auch Adverbien und Adjektive dazu gehören, da einige Verben in Partizipform (z.B. *helmeted*) als Adjektiv verwendet werden (*„Helmeted head of Athena facing.“* (ID: 138)) bzw. besonders im Deutschen als Adverbien erkannt werden(*„Hirschkuh nach rechts stehend“* (ID:4895)). Hinsichtlich anderer Entitätenarten können fehlerhafte Annotationen durch das Prüfen der Part-of-Speech Tags ebenfalls vermieden werden. Es sollen keine Pronomen, Determintative (Artikel), Adverbien, Adjektive, Verben und Hilfsverben annotiert werden, da diese Wortarten nicht als Einzelwort in den Entitätslisten vorkommen, welche hauptsächlich aus Substantiven bestehen. Lediglich in Phrasen könnnen diese Worttypen vorkommen (*Alexander der Große*), weshalb bei der Analyse von Phrasen keine Prüfung der Wortarten erfolgt. Allerdings ist das Identifizieren des Part-of-Speech Tags nicht fehlerfrei. Beispielsweise werden die Nomen *Tücher, Raben* und *Weizenähre* als Adverbien identifiziert. Dieser Fehler behindert bei diesen Worten die korrekte Annotation.

Zusätzlich zu der Suche nach Entitäten auf dem NLP-analysierten Satz, wird auch auf dem unveränderten Satz nach Entitäten gesucht. Ein direktes gefundes Match muss hier nicht weiter geprüft werden, da keines der Wörter angepasst wurde und wird immer als korrekte Annotation gewertet. Deshalb werden auch keine Part-of-Speech Tags ermittelt und geprüft. Es gibt Wörter, die unverändert gefunden werden können, allerdings durch fehlerhafte Part-of-Speech Zuweisung aussortiert werden. So wird *Weizenähre* als Verb und *Trophäe* als Adverb erkannt, obwohl beide Nomen sind. Durch das Ermitteln der Annotationen auf unveränderten Sätzen, und die direkte Übernahme der Übereinstimmungen in die Annotation, haben Fehler des Part-of-Speech Tags keine Auswirkugnen mehr.

Nachdem Annotationen mithilfe von NLP-Tools und auf dem unveränderten Satz erstellt sind, werden die getrennten Annotationen zusammengeführt und nach den Startkoordinaten der Einträge sortiert. Im Falle von Phrasen wie *palm branches* kann es vorkommen, dass Teilworte ebenfalls als einzelne Entitäten identifiziert werden. Deshalb wird bei diesen Überlappungen diejenige Annotation mit der größten Spannweite gewählt und Annotationen innerhalb dessen entfernt.
Im Fall der Komposita kommt es vor, dass mehrere Einträge für exakt dieselben Koordinaten existieren, die sich nur durch die Prüfzahl unterscheiden. Hier werden die Prüfziffern miteinander verglichen und diejenige Annotation gewählt, die die größte Prüfzahl hat.
Nach der Entfernung der Überlappungen werden die Prüfziffern entfernt und die Annotationen

nach ihren Koordianten aufsteigend sortiert ausgegeben. Die Annotation eines *Designs* ist nun beendet. Derselbe Prozess wird auf alle *Designs* angewandt.
Der Beispielsatz *„Kaiser (Caracalla) auf Pferd nach rechts, vor ihm Tyche mit Asklepiosstatuette im rechten Arm.“* (ID: 5645) besitzt nun folgende fertige Annotation:

[(0, 6, PERSON), (8, 17, PERSON), (23, 28, ANIMAL), (50, 55, PERSON), (60, 78, OBJECT)].

Mithilfe der Prüfziffer wird entschieden, dass der Entitätstyp von *Statuette* Vorrang hat in der Analyse des Kompositums *Asklepiosstatuette*.
Nachdem alle Sätze annotiert sind, werden diese abgespeichert und können zu einem späteren Zeitpunkt eingelesen und weiterverarbeitet werden. Ein Vorteil der analogen Struktur unserer Annotations stellt die Weiterverwendbarkeit der Annotation im Programmablauf von Deligio und Gencer (2021) dar.

4 Durchführung und Ergebnisse

In diesem Abschnitt werden die Ergebnisse der Annotationen der Datensätze mit den verschiedenen NLP-Methoden besprochen. Die gefundenen Entitäten der verschiedenen Annotationen werden miteinander und mit den Annotationen von Deligio und Gencer (2021) verglichen. Die Annotation von Deligio und Gencer (2021) wird in diesem Abschnitt als *manuelle Annotation* bezeichnet. Zudem werden die vorkommenden Fehlertypen und deren mögliche Behebung besprochen. Abschließend werden die ausgewählten Algorithmen der NLP-Methoden mit anderen Algorithmen dafür verglichen.

4.1 Durchführung

Die *Designs* werden in Deutsch und Englisch dreimal annotiert. Einmal mit Lemmatisierung als NLP-Methode, einmal mit Stemming, und einmal mit beiden Methoden, wobei hier separat eine Annotation mittels Lemmatisierung und eine mittels Stemming erstellt wurde und diese dann vermischt werden. Als Lemmatisierungsalgorithmus für Deutsch und Englisch wird die Lemmatisierungsfunktion von *SpaCy* genutzt. Die dadurch entstandenen Annotationen werden in diesem Abschnitt als *Lemma-Annotationen* bezeichnet. Stemming wird im Englischen mit dem *Snowball*-Stemmer von NLTK durchgeführt, und wird und im Deutschen mit *Cistem*. Diese Annotation wird als *Stem-Annotation* und die kombinierte Annotation der beiden Methoden wird als *Kombi-Annotation* bezeichnet. Die *Kombi-Annotation* enthält alle Entitäten, die in der *Stem-Annotation* oder *Lemma-Annotation* vorkommen. Aus diesem Grund wird diese Annotation genutzt, um die Gesamtleistung der Verwendung von NLP-Methoden zu beurteilen.

Die *Lemma-Annotationen* haben sprachunabhängig ca. 6 Minuten benötigt und die *Stem-Anno-*

tationen ca. 40 Sekunden. Die *Kombi-Annotationen* benötigen entsprechend ca. 6 Minuten 40 Sekunden.

4.1.1 Analyse der deutschen Annotationen

Tabelle 5 zeigt die Anzahl der insgesamt gefundenen Entitäten und den darin enthaltenene einzigartigen Entitäten in den deutschen Designs:

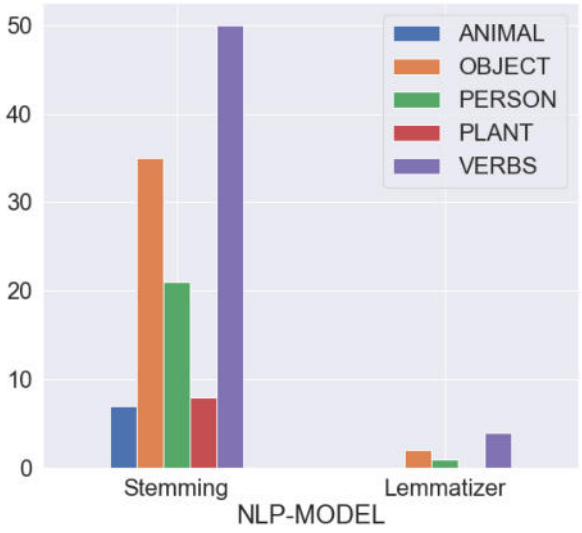
Tabelle 5: Vergleich der gefundenen Entitäten in den deutschen Designs

	Gesamtzahl annotierter Entitäten	**Annotierte einzigartige Entitäten**
Annotation aus Deligio und Gencer (2021) (*manuelle Annotation*)	26072	606
Annotation mit beiden Methoden (*Kombi-Annotationen*)	27171	803
Annotation mittels Stemming (*Stem-Annotationen*)	26992	800
Annotation mittels Lemmatisierung (*Lemma-Annotationen*)	26464	686

Man sieht große Unterschiede zwischen der *manuellen Annotation* und den Annotationen, die NLP-Methoden nutzen. Mithilfe der NLP-Methoden wurden insgesamt 1099 mehr Worte annotiert, von denen 201 einzigartige Entitäten sind, die bisher nicht von Deligio und Gencer (2021) identifiziert wurden. Dies zeigt, dass einige Wortformen bei der manuellen Eintragung in die Entitätenlisten übersehen wurden. Dieses Ergebnis entspricht den Erwartungen, da das Deutsche eine große Vielfalt an Flektionsformen besitzt. Eine genaue Analyse der neu gefundenen Entitäten findet sich in Abschnitt 4.3. Die meisten dieser neuen Entitäten sind durch den Stemmer entdeckt worden.

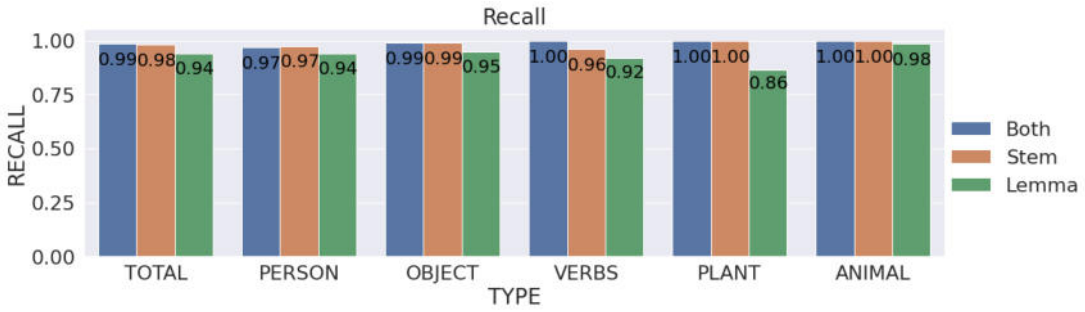
Abbildung 4 zeigt, aufgeteilt nach Entitätenkategorie, wie viele Entitäten ausschließlich durch den Stemmer gefunden wurden, und wie viele ausschließlich durch den Lemmatisierer gefunden wurden.

Abbildung 4: *Vergleich der Ergebnisse durch Lemmatisierung und Stemming (Deutsch)*



In allen Kategorien wurde durch Einsatz des Stemmer mehr Entitäten entdeckt als durch den Lemmatisierer. Grund dafür ist, dass der verwendete Lemmatisierer von *SpaCy* nur die Worte der *Designs* in einer Tabelle sucht und das zugeordnete Lemma ausgibt. Da die *Designs* des Corpus Nummorum allerdings viele unübliche Wortformen und Fremdworte enthalten wie*„Erymanthischen Eber“* und *„geschultertem“*, sind diese nicht in der Lookup Tabelle eingetragen. Der Stemmer hingegen entfernt lediglich Suffixe und Präfixe und ist deshalb in der Lage die vorliegenden Wörter auf ihren Stamm zurückzuführen.
Besonders viele Verben und Objekte werden von der *Stem-Annotation* annotiert, da von diesen eine größere Anzahl in den *Designs* vorkommt und diese in vielen flektierten Formen auftreten, die vom Stemmingalgorithmus auf den korrekten Stamm reduziert werden.

Abbildung 5: *Ergebnis der Berechnung des Recallwertes zwischen den Modellen und den manuellen Annotationen der deutschen Daten*

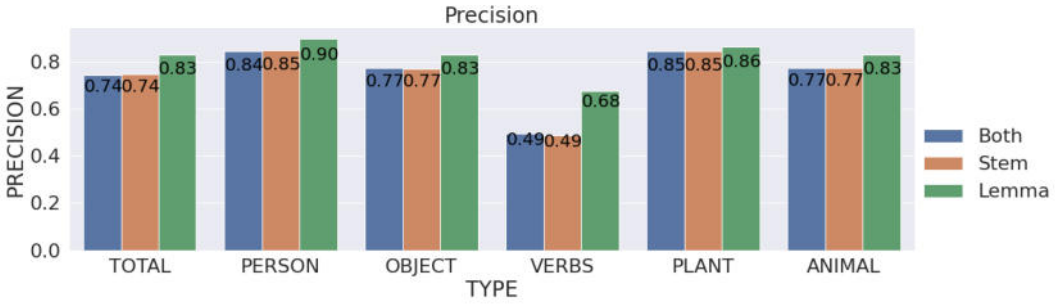


Wie bereits in Kapitel 2.4 besprochen, steht der Recall bei dem Vergleich der *Kombi-Annotation* mit der *manuellen Annotation* dafür, welcher Anteil der Entitäten von *manuelle Annotation* von den neuen Annotationen auf dieselbe Weise annotiert wird. Abbildung 5 zeigt den errechneten Recall in allen Entitätenklassen für *Kombi-Annotation, Lemma-Annotation* und *Stem-Annotation* an, sowie den Recallwert in allen Kategorien zusammen (TOTAL). Es ist zu erkennen, dass der Recallwert der *Stem-Annotation* in allen Kategorien über 95% liegt. Die *Lemma-Annotation* zeigt leicht schlechtere Werte, mit dem niedrigsten bei 86% für die Entitätenkategorie PLANT. Die *Kombi-Annotation* erreicht Recallwerte zwischen 97% und 100% . Folglich werden nahezu alle Entitäten der *manuellen Annotationen* von den neuen Annotationen gleich

analysiert. In Abschnitt 4.2 werden diejenigen Entitäten, die ausschließlich von der *manuellen Annotation* identifiziert wurden, genauer untersucht.

Wie in Kapitel 2.4 eingeführt steht Precision beim Vergleich der *Kombi-Annotation* mit der *manuellen Annotation* dafür, welcher Anteil der Entitäten der neuen Annotationen von der *manuellen Annotation* überhaupt nicht oder anders annotiert werden. Ein höherer Wert bedeutet hier also, dass weniger potenziell neue Entitäten identifiziert wurden, und ein niedriger Wert besagt, dass mehr potenzielle Entitäten entdeckt wurden. Da die *manuelle Annotation* keinen Goldstandard für die Annotation der *Designs* darstellt, ist es möglich, dass die neuen Entitäten korrekt annotiert wurden. Dies wird durch manuelle Prüfung der Ergebnisse bestimmt. In Abschnitt 4.3 werden diese neuen Entitäten besprochen.
Abbildung 6 stellt die errechneten Precisionwerte in allen Entitätenklassen für beide NLP-Methoden, Stemming und Lemmatisierung dar.

Abbildung 6: *Ergebnis der Berechnung des Precisionwertes zwischen den Modellen und den manuellen Annotationen der deutschen Daten*



Der Precisionwert in allen Entitätenkategorien (TOTAL) liegt bei der Annotation mit beiden Methoden bei 75%, wobei die *Lemma-Annotation* höhere oder gleiche Werte in allen Kategorien aufwiest, wie die *Stem-Annotation*. In der Kategorie VERBS hat die *Lemma-Annotation* einen besonders hohen Wert (68%), wohingegen die *Stem-Annotation* und die *Kombi-Annotation* 49% erreicht. Das bedeutet, dass diese beiden Annotationen äußerst viele neue Verbformen identifiziert haben, was zu dem hohen Grad an möglichen Konjugationsformen im Deutschen passt.

4.1.2 Analyse der englischen Annotationen

Tabelle 6 zeigt die Anzahl der annotierten Wörter für Englisch in den verschiedenen Annotationen. Im Englischen treten, wie man an der *Kombi-Annotation* sieht, insgesamt 802 neue Annotationen auf, von denen 89 neue Entitäten sind. Die *Stem-Annotation* hat, wie im Deutschen, auch hier mehr Entitäten entdeckt, als die *Lemma-Annotation*. Da Englisch weniger Flektionsformen besitzt, ist es nicht verwunderlich, dass insgesamt weniger neue Entitäten entdeckt wurden als bei der deutschen Annotation.

Tabelle 6: Vergleich der gefundenen Entitäten in den englischen Designs

	Gesamtzahl annotierter Entitäten	**Annotierte einzigartige Entitäten**
Annotation aus Deligio und Gencer (2021)	30069	731
Annotation mit beiden Methoden (*Kombi-Annotation*)	30871	820
Annotation mithilfe von Stemming (*Stem-Annotation*)	30824	818
Annotation mithilfe von Lemmatisierung (*Lemma-Annotation*)	30759	796

Abbildung 7 zeigt aufgeteilt nach Entitätenkategorien, wie viele Entitäten ausschließlich mittels Stemming gefunden wurden, und wie viele ausschließlich mittels Lemmatisierer gefunden wurden.

Durch Stemming werden mehr Objekte, Pflanzen und Personen identifiziert, aber weniger Verben als durch Lemmatisierung. Der Unterschied in der Anzahl der gefundenen Wörter ist insgesamt weniger groß als im Deutschen. Grund dafür ist auch hier die geringere Anzahl an Flektionsformen. Da weniger Suffixe und Suffixkombinationen existieren, ist es leichter verschiedene Wortformen auf dasselbe Lemma oder denselben Stamm zurückzuführen.

Abbildung 7: *Vergleich der Ergebnisse durch Lemmatisierung und Stemming (Englisch)*

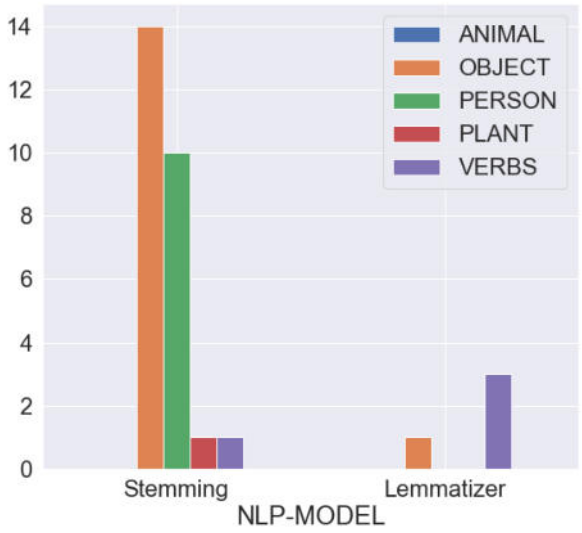


Abbildung 8 zeigt den errechneten Recall in allen Entitätenklassen für die *Kombi-Annotation, Stem-Annotation* und *Lemma-Annotation*. Es ist zu erkennen, dass der Recall für Stemming bei mindestesn 99% liegt, außer in der Kategorie VERBS, in der 96% Recall erreicht werden. Mit Lemmatisierung wird ein Recall zwischen 100% und 98% erreicht, wobei nur bei den Verben mithilfe von Lemmatisierung ein höherer Wert erreicht wird, als durch Stemming. Die *Kombi-Annotation* erreicht ebenfalls einen Wert von 99% über alle Kategorien (TOTAL). Aus diesen Beobachtungen lässt sich schließen, dass fast alle Entitäten der *manuellen Annotation* von der neuen *Kombi-Annotation* gefunden werden.

Abbildung 8: *Ergebnis der Berechnung des Recallwertes zwischen den Modellen und den manuellen Annotationen der englischen Daten*

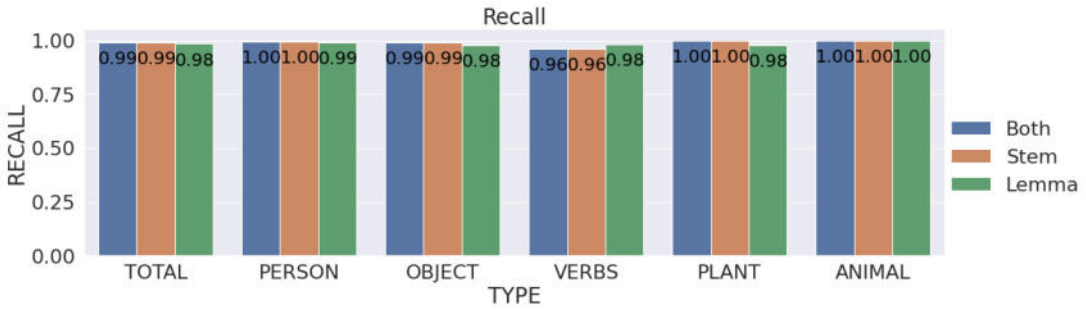
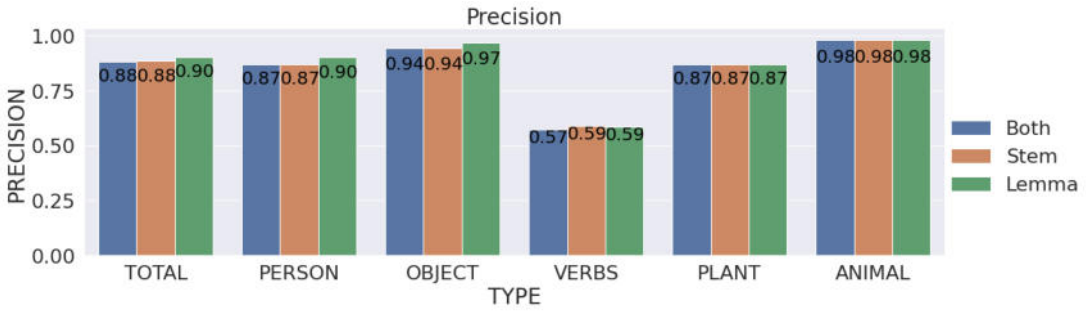


Abbildung 9 zeigt die errechneten Precisionwerte in allen Entitätenklassen für *Kombi-Annotation, Stem-Annotation* und *Lemma-Annotation* an. Der Precisionwert über alle Entitätenkategorien (TOTAL) liegt bei 88% für die *Kombi-Annotation*. Dieser Wert ist etwas höher als der Wert bei den deutschen *Designs* (75%). Aber ähnlich wie im Deutschen erzielt der Lemmatisierer entweder den gleichen, oder einen höheren Wert, als der Stemmer in allen Kategorien. Ebenfalls sind die meisten neuen Entitäten hier aus der Klasse VERBS.

Abbildung 9: *Ergebnis der Berechnung des Precisionwertes zwischen den Modellen und den manuellen Annotationen der englischen Daten*



Diese Ergebnisse zeigen, dass trotz der geringeren Anzahl an Flektionsformen auch die Annotation der englischen *Designs* vom Einsatz der NLP-Methoden profitiert, wenn auch weniger als die der deutschen *Designs*. Der Einsatz des Lemmatisierers hat weniger neue Entitäten generiert, da auch hier zusätzlich zu den Regeln eine Lookup Tabelle eingesetzt wird, mit denen geprüft wird, ob das erzeugte Lemma eines Wortes ein echtes Wort ist. Da die Entitätenlisten aber viele griechische Fremdworte enthalten, werden diese nicht als Lemma erkannt. Der Einsatz von Stemming bietet, wie im Deutschen, den Vorteil, dass auch Fremdworte auf einen Stamm zurückgeführt werden können, und dass der Erfolg nicht von Lookup Tabellen abhängig ist.

4.2 Besprechung der nicht entdeckten Entitäten

Dieser Abschnitt untersucht diejenigen Entitäten, die in der *manuellen Annotation* identifiziert wurden, aber nicht von der *Kombi-Annotation*, die die Entitäten enthält, die durch Stemming oder Lemmatisieren identifiziert wurden. Zu den neuen Entitäten fallen auch Begriffe, denen ein anderer Entitätentyp zugeordnet wird. Die Tabellen im Anhang zeigen die vollständige Liste der Entitäten, die in diese Kategorie gehören. Für die deutschen Annotationen ist es *Unterabschnitt A.3*, und für die Englischen *Unterabschnitt A.4*.
Es existieren vier verschiedene Gründe für das Zuordnen eines anderen Entitätentyps oder Fehlen dieser Begriffe in der *Kombi-Annotation*:

- Bei den Worten *Alexander des Großen, Stadtgöttinnen* und *Standarten* liegt der Fehler bei den NLP-Methoden. Die Phrasen in den *Designs* und der eigentlich dazugehörigen Eintrag in der Entitätenliste werden nicht auf dasselbe Lemma oder denselben Stamm zurückgeführt. Tabelle 7 zeigt, welche Formen einer Entität welches Ergebnis durch Lemmatisierung und Stemming erhalten. Man sieht, dass sowohl die erzeugten Lemmas, als auch die Stämme nicht übereinstimmen.

Tabelle 7: Lemma und Stemming

Original	**Lemma**	**Stem**
Alexander des Großen Alexander der Große	Alexander der großen Alexander der große	alexa des gross alexa der gross
Stadtgöttinnen Stadtgöttin	Stadtgöttinnen Stadtgöttin	statdgottinn stadtgotti
Standarten Standarte	Standarten Standarte	Standart Standar

- Das Partizip *Sailing* wird lediglich von der *manuellen Annotation* als Verb annotiert. Durch die NLP-Methoden wird das Suffix *-ing* entfernt und dieselbe Form wie das Objekt *Sail* erzeugt. Außerdem wird dieses Wort vom Part-of-Speech Tagger als Nomen identifiziert. Aus diesen Gründen wird *Sailing* als OBJECT annotiert. Das Eintragen von *Sailing* in die Entitätenliste VERBS ist eine Möglichkeit dies zu umgehen. Der vorgestellte Algorithmus zur Annotation nutzt auch ein direktes Matching auf den unveränderten Wörtern, wodurch das Ergebnis der Lemmatisierung oder dem Stemming umgangen wird und somit diese Verbformen als Entitäten der Klasse VERBS erkannt werden. Bei anderen Verben mit dem Suffix *-ing* ist dieser Fehler nicht aufgetreten, da entweder der erzeugte Stamm nicht in einer Entitätenliste eingetragen ist, oder es als Verbform erkannt wurde.

- Einige der Entitäten, die mittels der NLP-Methoden nicht entdeckt wurden, sind in der *manuellen Annotation* falsch annotiert. Dort wird in diesen Fällen nicht die gesamte Phrase genutzt, sondern nur einen Teil davon, wohingegen die Annotation in dieser Arbeit die vollständige Phrase identifiziert hat, wie in Tabelle 8 gezeigt wird.

Tabelle 8: Vergleich der Teilphrasen mit den korrekten, vollständigen Phrasen

Teilphrase der *manuellen Annotation*	**Gesamtphrase in der Entitätenliste und den *Designs***
Kotys IV	Kotys IV.
Gordian III	Gordian III.
Faustina I	Faustina I.
Hades	Hades-Serapis
leaned	leaned on
Seleucus I	Seleucus I Nikator
Cista/cista	Cista/cista mystica

- Der letzte Fall betrifft lediglich as Wort *beehive*. Dieses ist sowohl in der Entitätenliste von OBJECT, als auch von ANIMAL enthalten. Der hier beschriebene Algorithmus weist dem Wort allerdings nur einen Entitätenbezeichner zu. Bei einem anderen Bezeichner als dem, der von der *manuellen Annotation* zugewiesen wird, wird dies deshalb als Unterschied in der Annotation erkannt. Eine Entfernung der doppelten Einträge aus den Entitätenlisten behebt dieses Problem.

4.3 Besprechung der neu entdeckten Entitäten

Dieser Abschnitt untersucht diejenigen Entitäten, welche in der *Kombi-Annotation* identifiziert wurden, aber nicht in der *manuellen Annotation*. Zuerst werden die Entitäten besprochen, welche korrekt identifiziert wurden, sowie bestimmt, warum diese Wörter mithilfe der NLP-Methoden entdeckt wurden. Dann werden die Entitäten, die falsch annotiert wurden, besprochen und beschrieben, wie diese Fehler zustande kommen. Eine Übersicht aller ausschließlich von der *Kombi-Annotation* gefundenen Entitäten findet sich im Anhang - für Deutsch in *Unterabschnitt A.2* und für Englisch in *Unterabschnitt A.1*.

Korrekt annotierte neue Entitäten

Nach manueller Überprüfung der 205 neu identifizierten Entitäten im Deutschen, sind 199 davon korrekt ermittelt worden, was 97% ausmacht. In den englischen *Designs* sind insgesamt 96 neue Entitäten gefunden worden, 92% richtig annotiert sind. Nach dem Vergleich der neu gefundenen Entitäten mit den Annotationen und Entitätenlisten von Deligio und Gencer (2021), können drei verschiedene Gründe angegeben werden, warum diese Worte mittels der NLP-Methoden annotiert werden, aber in der *manuellen Annotation* nicht.

- Die erste Kategorie von Worten enthält flektierte Wortformen, die nicht in den Entitätenlisten von Deligio und Gencer (2021) eingetragen sind z.B. *Kretischen Stiers*, *liegendem* und *feeds*. Die meisten neu annotierten verben gehören in diese Kategorie. Zusätzlich fallen, besonders im Deutschen, Objekte und Tiere darunter. Grund dafür ist wieder die Vielfalt der Flektionsformen in dieser Sprache. Dazu gehören auch die Partizip Perfekt Formen *Veiled, Helmeted, Diademed, Cuirassed* und *Garlanded*. Diese wurden als For-

men der Objekte *Veil, Helm, Diadem, Cuirass* und *Garland* erkannt und bedeuten in diesem Kontext, dass die Objekte von einer Person getragen werden.

- Die zweite Kategorie enthält die Komposita. Im Deutschen sind 22 neue Komposita entdeckt worden:
  Krabbenscheren, Asklepios-Schlange, Säulenkapitell, Reiterstatue, Turmkrone, Asklepios-Statuette, , Linienkreuz, Säulenbasis, Aslepiosstatue, asklepiosstatuette, Bandschleifen, Helmbusch, Quellgefäss, Adlerkopf, Göttinnenkopf, Schlangenkopf, Dionysos-kind, Plutos-kind, Lanzenspitze, Panther-Biga und *Säulenmonument*.
  Einige von diesen sind ebenfalls in verschiedenen flektierten Formen vorgekommen. Im Englischen wurde nur ein Kompositum, *horseback*, entdeckt und als ANIMAL annotiert. Dies entspricht der Erwartung, dass Komposita im Englischen unüblich sind. Wie von Deligio und Gencer (2021) beschrieben, stellten diese Wörter eine Herausforderung für die *manuelle Annotation* dar, da bei dieser Teilworte nicht erkannt werden.

- Die dritte Kategorie enthält Phrasen aus mehreren Worten, von denen nicht die vollständige Phrase in der *manuellen Annotation* erkannt wurde. Bei *Athena Promachos* und *Tyche Poleos* beispielsweise wird nur der erste Name erkannt, wohingegen der hier vorgestellte Algorithmus die vollständige Phrase annotiert.

Falsch annotierte, neue Entitäten

In der *Kombi-Annotation* wurden in den deutschen *Designs* lediglich 6 von 205 Entitäten,also nur 3%, falsch annotiert, die nicht von der *manuellen Annotation* gefunden wurden, und bei den Englischen 8 von 96 Entitäten, also 8%.
Diese lassen sich in 5 verschiedene Kategorien einteilen.

- Die erste Kategorie enthält Wörter, die auf denselben Stamm bzw. dasselbe Lemma zurückgeführt werden, wie ein komplett anderes Wort. Sowohl im Deutschen als auch im Englischen werden *Hermes*, *herm* und *herme* auf dieselbe Grundform zurückgeführt. Allerdings ist *Hermes* der Name einer Person und *herm* ein Objekt. In diesem Fall wird also einem Wort der falsche Entitätenbezeichner zugeordnet. Fehler dieser Art können nur durch einen verbesserten Lemmatisierer oder Stemmer gelöst werden.

- Die zweite Kategorie beinhaltet Wörter, denen ein falscher Part-of-Speech Tag zugewiesen wird. Einzelne Wörter können nur als Nomen oder Verben eine Entität sein, da andere Wortarten lediglich innerhalb von Phrasen verwendet werden. Einigen Einzelwörtern wird allerdings die falsche Wortart vom Part-of-Speech Tagger zugewiesen. Die Adjektive *Jugendlichen* und *Youthful* sowie die Partizipien *Stepping, sailing* und *crossing* werden als Nomen getaggt und somit annotiert. Da die flektierten Formen dieser Worte auch nominal sein können, ist es vom satzkontext der Worte abhängig, ob der Tagger hier einen Fehler gemacht hat. Nach manueller Prüfung der Sätze, in denen diese Worte falsch getaggt wurden, ist bestätigt, dass diese Worte dort als Adjektive und Verben verwendet werden.

- Die dritte Kategorie betrifft Wörter, die fälschlicherweise als Kompositum erkannt werden. Hier ist es nur ein Wort *schreitender*. Dieses wird in die Teilworte *schreiten-* und *-der* aufgespalten und *schreiten-* wird auf denselben Stamm wie *Schrein* zurückgeführt. Aus diesem Grund wird diese Verbform als OBJECT annotiert. Der Fehler ist eindeutig auf die verwendete Library zum Erkennen von Komposita zurückzuführen.

- Die vierte Kategorie beinhaltet Komposita, bei denen das wesentliche, sinngebende Teilwort nicht in den Entitätenlisten vorhanden ist. Dies betrifft die Komposita *Strymon-Gegend* und *Elefantenzahn*. Weder *Gegend* noch *Zahn* sind als Entitäten eingetragen, weshalb die genannten Worte als PERSON und ANIMAL annotiert werden. Fehler dieser Art können durch Einfügen der Teilworte in die Entitätenliste behoben werden.

- Die letzte Kategorie besteht aus Fehlern in den *Designs*. In mindestens zwei *Designs* wurden Leerzeichen vergessen und das dadurch entstandene Wort als Kompositum identifiziert, wie *darüberAdler* und *ihrPferd*. Dementsprechend liegt dieser Fehler nicht beim Algorithmus.

4.4 Vergleich mit anderen Lemmatisierungs- und Stemmingalgorithmen

In diesem Abschnitt wird untersucht, ob andere Lemmatisierungs- und Stemmingalgorithmen bessere Ergebnisse erzielen können, als die bisher verwendeten:
Für das Deutsche wird zur Erstellung der *Lemma-Annotation* der Algorithmus von *SpaCy* und zur Erstellung der *Stem-Annotation* der Algorithmus von *Cistem* eingesetzt. Für das Englische wird ebenfalls die Lemmatisierungsfunktion von *SpaCy* eingesetzt, und *Snowball* zur Erstellung der *Stem-Annotation*. Für jeden dieser Algorithmen wird in diesem Abschnitt eine Alternative ausgewählt und in Kombination miteinander angewendet. Die daraus entstandenen Annotationen werden dann mit der *Kombi-Annotation* der jeweiligen Sprache verglichen.

Deutsche Annotationen

Für das Deutsche wird als alternative Lemmatisierungsfunktion der HannoverTagger (*HanTa*) verwendet. Als neuer Stemmingalgorithmus wird *Pystem* eingesetzt, der eine Implementierung von *Snowball* für Deutsch darstellt. Der in Kapitel 3 vorgestellte Programmablauf wird nicht verändert, sondern lediglich der verwendete Lemmatisierer und Stemmer angepasst. Insgesamt werden drei neue Annotationen erstellt:

- Lemmatisierungsalgorithmus von *SpaCy* und Stemmingalgorithmus von *Pystem* (*SpaCy + Pystem*)
- Lemmatisierungsalgorithmus von *HanTa* und Stemmingalgorithmus von *Cistem* (*HanTa + Cistem*)
- Lemmatisierungsalgorithmus von *HanTa* und Stemmingalgorithmus von *Pystem* (*HanTa + Pystem*)

Für die Ergebnisse dieser werden Recall und Precision mit Vergleich zu der *Kombi-Annotation* mit *SpaCy* und *Cistem* berechnet und nach Entitätenkategorien aufgespalten.

Abbildung 10: *Vergleich des Recalls der Ergebnisse der NLP-Methoden (Deutsch)*

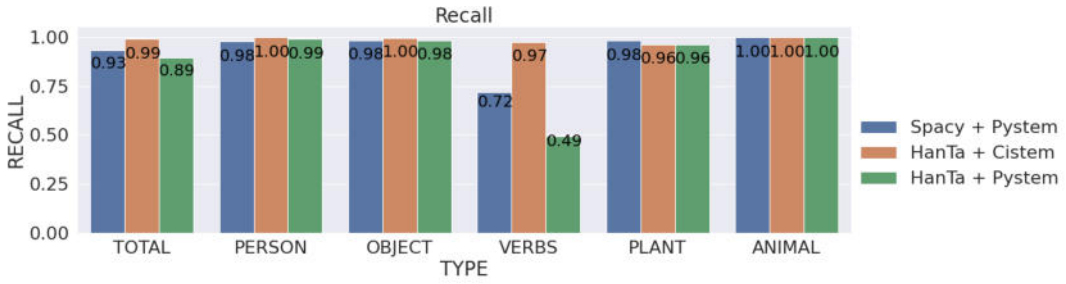
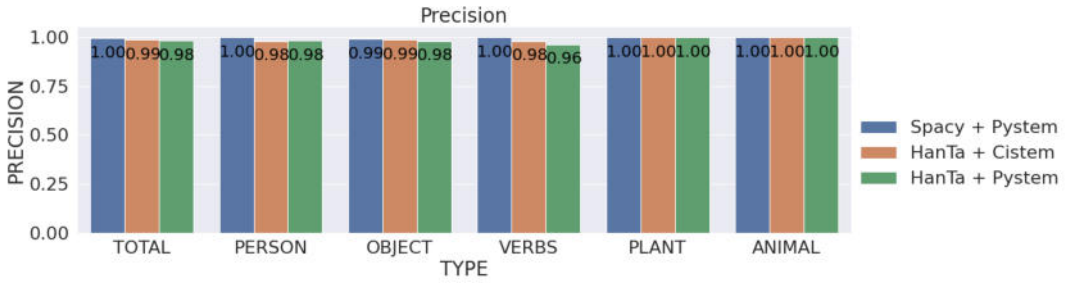


Abbildung 10 zeigt die Ergebnisse der Berechnung des Recalls der Annotationen der drei Kombinationen im Vergleich zur *Kombi-Annotation*. Die Kombination *SpaCy + Pystem* erzielt nur in der Kategorie ANIMAL 100% Recall. Die Kategorie VERBS enthält den niedrigsten Recallwert. In allen Kategorien zusammen (TOTAL) erreicht diese Kombination den Wert 93%. Die Kombination *HanTa + Cistem* hat in allen Kategorien den höchsten Recallwert erreicht. Im Falle von VERBS ist es die einzige Kombination über 95%. Die Kombination *HanTa + Pystem* erzeugt die schlechtesten Ergebnisse mit 89% in TOTAL. In der Kategorie VERBS werden damit nur 49% der Entitäten der *Kombi-Annotation* gefunden.

Abbildung 11 zeigt die erreichten Precisionwerte der drei Kombinationen im Vergleich zur *Kombi-Annotation*. Bei der Betrachtung der Precisionwerte fällt auf, dass jeder Wert größer als 95% ist. Das bedeutet, dass nur äußerst wenige neue Entitäten von all diesen Kombinationen annotiert werden.

Abbildung 11: *Vergleich der Precision der Ergebnisse der NLP-Methoden (Deutsch)*



Im nächsten Schritt werden die neu gefundenen Entitäten, sowie die nicht gefundenen Entitäten manuell daraufhin geprüft, ob sie korrekt annotiert werden. Im Anhang unter *Unterabschnitt A.6* sind alle Entitäten gesammelt, die ausschließlich von den jeweiligen Kombinationen und ausschließlich von der *Kombi-Annotation* entdeckt wurden. *SpaCy + Pystem* hat insgesamt drei

neue Entitäten annotiert, und alle drei sind korrekt annotiert. Allerdings sind 94% d.h. 51 von 54 nicht gefundenen Entitäten fälschlicherweise nicht gefunden worden.
HanTa + Cistem hat 12 neue annotierte Entitäten, aber von diesen sind nur ein Drittel korrekt annotiert. Allerdings konnten hiermit zwei neue Tippfehler in den *Designs* gefunden werden, *„Lorbeerkranzin“* und *„Felsten“*. Nur 7 Entitäten sind von dieser Kombination nicht annotiert worden, allerdings sind 6 davon fälschlicherweise nicht annotiert.
HanTa + Pystem hat von 13 neuen Annotationen nur 6 korrekt annotiert, und von 86 nicht gefundenen Entitäten sind 83 fälschlicherweise nicht annotiert worden.

Diese Ergebnisse zeigen, dass die Verwendung des *Pystem* Stemmers keinen Vorteil gebracht hat, sondern besonders die Quantität der gefundenen Verben im Deutschen verringert. Der *HanTa* Lemmmatisierer hingegen erreicht 99% Recall und Precision im Vergleich zur *Kombi-Annotation* bei gleichem Stemmingalgorithmus und somit ungefähr dieselben Ergebnisse erbringt, wie der Lemmatisierer von *SpaCy*.

Englische Annotationen
Als alternative Lemmatisierungsfunktion wird *Wordnet* genutzt, welches von NLTK zur Verfügung gestellt wird. Als Alternative zu *Snowball* wird zum Stemming der *Krovetz-Stemmer* eingesetzt. Wie bei den deutschen Annotationen, werden auch hier drei neue Annotationen erstellt, welche die Kombinationsmöglichkeiten der Algorithmen abdecken.

- Lemmatisierungsalgorithmus von *Wordnet* und Stemmingalgorithmus von *Snowball* (*Wordnet + Snowball*)
- Lemmatisierungsalgorithmus von *SpaCy* und Stemmingalgorithmus von *Krovetz* (*SpaCy + Krovetz*)
- Lemmatisierungsalgorithmus von *Wordnet* und Stemmingalgorithmus von *Krovetz* (*Wordnet + Krovetz*)

Die Ergebnisse werden dann durch Berechnung von Recall und Precision mit der *Kombi-Annotation* unter Verwendung von *SpaCy* und *Snowball* verglichen .

Abbildung 12: *Vergleich des Recalls der Ergebnisse der NLP-Methoden (Englisch)*

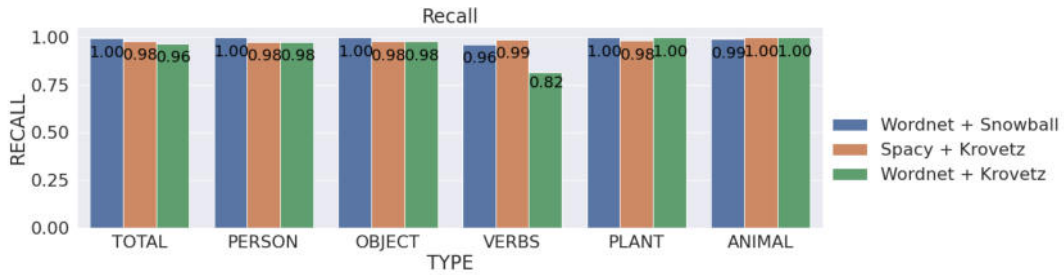


Abbildung 12 zeigt die errechneten Recallwerte der Kombinationen. Der Recallwert von al-

len Kombinationen in allen Kategorien liegt über 98% außer in der Kategorie VERBS. Dort liegt der Recall von *SpaCy + Krovetz* nur bei 82%, und von *Wordnet + Snowball* bei 96%. Die Kombination *Wordnet + Krovetz* zeigt in allen Kategorien zusammen (TOTAL) den niedrigsten Recall mit 96%. Die Kombination *Wordnet + Snowball* zeigt in 3 von 5 Entitätenkategorien jeweils 100% Recall. Das bedeutet, dass alle Entitäten dieser Kategorien, die von der *Kombi-Annotation* gefunden werden, auch mit dem *Wordnet*-Lemmatisierer auf dieselbe Art annotiert werden.

Abbildung 13: *Vergleich der Precision der Ergebnisse der NLP-Methoden (Englisch)*

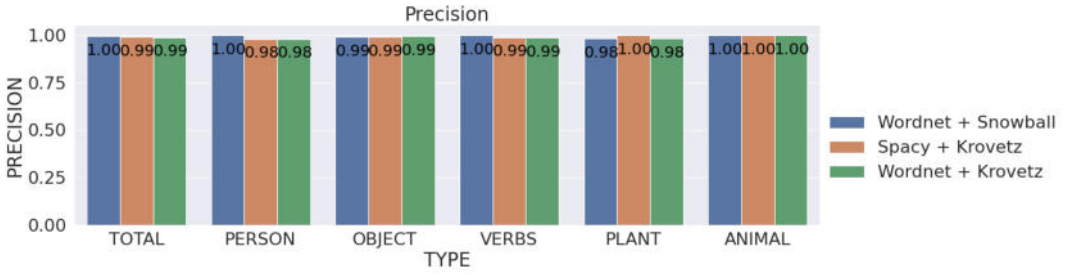


Abbildung 13 zeigt die errechneten Precisionwerte zwischen der *Kombi-Annotation* und den drei Annotationen der Kombinationen. Die Precisionwerte von allen drei Kombinationen in allen Entitätenkategorien liegen bei mindestens 98%. Es gibt kaum einen Unterschied in den Werten der Kombinationen untereinander. Das bedeutet, dass insgesamt nur sehr wenige weitere Entitäten annotiert wurden.

Auch hier werden die neu gefundenen Entitäten, sowie die nicht gefundenen Entitäten manuell daraufhin geprüft, ob sie korrekt annotiert werden. Im Anhang unter *Unterabschnitt A.5* sind alle Entitäten genannt, die ausschließlich von den jeweiligen Kombinationen und ausschließlich von der *Kombi-Annotation* entdeckt wurden.
Diese Ergebnisse der Betrachtung der Recall und Precisionwerte der Kombinationen mit der ursprünglichen NLP-Annotation, zeigen dass es zwischen den Lemmatisierungsalgorithmen von SpaCy und NLTK kaum einen Unterschied in der Qualität der Ergebnisse gibt. In der Kombination *Wordnet + Snowball* vier neue Annotationen entdeckt worden, von denen zwei korrekt sind. Darüber hinaus wurden vier Entitäten nicht gefunden, von denen drei fälschlicherweise nicht annotiert wurden. die anderen Kombinationen haben mehr Fehler erzeugt. Bei *SpaCy + Krovetz* sind nur zwei von neun neu gefundenen Annotationen korrekt und 11 von 16 nicht gefundenen Annotationen wurden fälschlicherweise nicht identifiziert. Bei *Wordnet + Krovetz* sind ebenfalls nur zwei von neun neuen Annotationen richtig identifiziert, aber 24 von 29 nicht gefundenen Annotationen sind fehlerhaft. Die Anzahl erfolgreicher Annotationen des *Krovetz Stemmer* liegt also deutlich unter der von *Snowball*.

Sowohl bei den deutschen, als auch den englischen *Designs* hat keiner der in diesem Abschnitt getesteten Lemmatisierer und Stemmer einen deutlichen Vorteil gegenüber den bisherigen Ergebnisses erbracht. Die Anzahl gefundener Entitäten hat sich kaum erhöht, sondern ist entweder auf gleichem Niveau oder hat sich verrignert. Besonders Verben bereiten den Stemmingalgorithmen *Krovetz* und *Pystem* Schwierigkeiten. Die getesteten Lemmatisierungsalgorithmen (*HanTa* und *Wordnet*) bieten gute Alternativen zu SpaCy.

5 Fazit und Ausblick

Das Ziel dieser Arbeit war die Vereinfachung der Annotation des Modells von Deligio und Gencer (2021), sodass auf die manuelle Eintragung aller Flektionsformen eines Eintrags in die dazugehörigen Entitätentabellen verzichtet werden kann. Dafür wurde im Verlauf dieser Arbeit eine Annotationsfunktion vorgestellt, die Lemmatisierung und Stemming verwendet. Durch den Einsatz dieser Methoden sollen die Flektionsformen von Wörtern auf eine gemeinsame Grundform zurückgeführt werden. Auch Komposita und Partizipformen werden dadurch direkt annotiert.

Die Ergebnisse zeigen, dass die so erzeugten Annotationen mit der Annotation von Deligio und Gencer (2021) vergleichbar sind. Die Annotationen der eingesetzten Stemmingalgorithmen (*Cistem* für Deutsch und *Snowball* für Englisch) erzielen einen höheren Recall und eine höhere Anzahl neu gefundener Entitäten als der Lemmatisierer von SpaCy (vgl. Kapitel 4.1.). Eine Kombination von Lemmatisierer und Stemmer ergibt die meisten korrekten Annotationen, mit einem Recall von 99% bei deutschen und englischen *Designs*, im Vergleich zur Annotation von Deligio und Gencer (2021). Darüber hinaus werden neue Entitäten identifiziert, besonders Verben, die bisher nicht annotiert wurden. Im Deutschen werden 201 neue Entitäten entdeckt, und im Englischen 89. Dies zeigt den besonderen Nutzen der NLP-Methoden in einer Sprache mit hohem Flektionsgrad wie das Deutsche.
Im Vergleich mit anderen Algorithmen der NLP-Methoden haben sich die Lemmatisierer *Wordnet* für Englisch und *HanTa* für Deutsch bewährt. Deren Ergebnisse beinhalten 99% derselben Annotationen wie der *SpaCy* Lemmatisierer (vgl. Kapitel 4.4) bei gleichem Stemmingalgorithmus. Der Vergleich der Stemmingalgorithmen *Cistem* für Deutsch und *Snowball* für Englisch zeigt jedoch, dass die alternativen Algorithmen *Snowball* für Deutsch und *Krovetz-Stemmer* für Englisch schlechtere Ergebnisse erbringen.

An verschiedenen Stellen hat das hier vorgestellte Programm Schwierigkeiten. Wie in Kapitel 4.2 und 4.3 beschrieben weist der Part-of-Speech Tagger nicht immer die korrekte Wortart zu, wobei besonders im Englischen Verben, die im Partizip Präsenz stehen davon betroffen sind. Diese werden deshalb als Nomen erkannt (vgl. *sailing, crossing*). Hinzu kommt, dass die Lemmatisierer und Stemmer in einigen Fällen ein falsches Grundwort ermitteln. Dies führt zu falschen Annotationen. Der Part-of-Speech Tagger und die Lemmatisierungsalgorithmen

entstammen allerdings der Library SpaCy, die regelmäßig weiterentwickelt wird, sodass diese Aspekte sich in Zukunft allein durch ein Upgrade auf die neuste Version von SpaCy verbessern könnten.

Ausgehend von den Ergebnissen dieser Arbeit aus wäre es interessant die Auswirkungen der größeren Menge an annotierten Entitäten auf die *Named Entity Recognition* und *Relation Extraction* von Deligio und Gencer (2021) zu untersuchen. Die Vergrößerung des Trainingsdatensatz des *NER*-Modells könnte in einer verbesserten *Entity Recognition* resultieren, sodass wesentlich mehr Entitäten entdeckt werden können. Dies kann auch mehr komplett neue Entitäten identifizieren, die nicht in den Entitätentabellen eingetragen sind.

Wenn auch Münzbeschreibungen in anderen Sprachen als Deutsch und Englisch annotiert werden sollen, dann müssen Lemmatisierungs- oder Stemmingfunktionen für diese Sprachen ermittelt werden und zu der in dieser Arbeit entwickelten Annotationsfunktion hinzugefügt werden. Während diesem Vorgang müssen die sprachlichen Unterschiede der neuen Sprache zum Deutschen und zum Englischen ermittelt werden, da dies Konsequenzen für die Mechaniken der Annotationsfunktion haben kann. Im Hinblick auf die Entitätenkategorien kann es bspw. notwendig sein, dass andere Filterkriterien definiert werden müssen, um Wörter nicht falsch zuzuordnen. Falls beispielsweise andere Wortarten als Adverb, Adjektiv und Verb Teil der Entitätenkategorie VERBS sein kann, dann muss dies hinzugefügt werden. Auch kann es sein, dass, je nach betrachteter Sprache, das Kopf-rechts-Prinzip nicht mehr greift. dann muss eine angemessenere Heuristik zur Bestimmung des sinntragenden Teilwortes genutzt werden

A Tabellarische Aufzählung der annotierten Entitäten

A.1 Übersicht der gefundenen Entitäten der Kombi-Annotation (Englisch)

Tabelle 9: Übersicht der gefundenen Entitäten der Kombi-Annotation (Englisch)

Kategorie	Korrekt annotiert	fehlerhafte annotiert
PERSON	Agrippina II, pan, Gordian III, moon, heroes, earth, nike, Charite, Seleucus I Nikator, lyra, Nymphs, corybant, genius, prisoners, Plutos, Tyche Poleos, Senate, Agrippina Minor, prisoner, victory, caracalla, Agrippina I, herm, youth, hero, Athena Promachos, Hero, Apollo Smintheus, Hermes Perpheraios, figur	youthful, Youthful, Herm,
OBJECT	Veiled, cubit rule, cista mystica, quadrigas, crown juwel, clothing, Cuirassed, kneeing, coiling, Cista mystica, garlanded, Mitras, Helmeted, Diademed	Kneeing, Sailing, crossing, Hermes, stepping,
PLANT	Grain ear, palm tree, branch, grain ear, ear of corn, Ear of corn, leafes	
ANIMAL	horseback	beehive
VERBS	receive, leaned on, extended, coiled, fly, curled, drawn, reclining on, formed, resting on, escorted by, advanced, set, held, riding on, stands, Coiled, creep, seated on, crossed, crowned, rests, covered, feeds, raise, advances, leaning on, Raised, hold, carry, rested on, holds, stand, raised, rests on	

A.2 Übersicht der gefundenen Entitäten der Kombi-Annotation (Deutsch)

Tabelle 10: Übersicht der gefundenen Entitäten der Kombi-Annotation (Deutsch)

Kategorie	**Korrekt annotiert**	**Fehlerhafte annotiert**
PERSON	Apollon Smintheus, Niken, Kotys IV., Soldaten, Faustina II., Agrippina II., Hermes Perpheraios, Athena Promachos, Flussgöttern, Tyche Poleos, Gaius Caesar, Agrippina I., Faustina I., Dioskuren, Silenen, Begleitern, Gordian III., Hero, Hades-Serapis, Kotys I., Gefangenem, Seleukos I., Kaisers, Philippus II., Heroen, Dionysoskind, Ptolemaios III Euergetes, Tanzende, Flußgott, Plutos-Kind	Fängen, jugendlichen, Strymon-Gegend, Jugendlichen, Herme
OBJECT	Kränzen, Turmkrone, Pfeiler, Pferdeprotomen, Giebeln, Reiterstatue, Herme, Palmzweigen, Bogens, Toren, Ruderer, Stadttores, Kranze,, Asklepiosstatuette, Türmen, Rammsporen, Linienkreuz, Cista mystica, Palmenzweig, Schlangenkopf, Bandschleifen, Füllhörnern, Bandschleife, Quellgefäss, Sternen, Köpfen, Knieen, Tempeln, Altars, Helmbusch, Knien, Tempels, amphora, Adler-Zepter, Amphore, Throns, Antefixe, Ornamenten, Gefäßen, Himations, LorbeerKranz, Palmzweige, Prorae, Mitras, Ohrringen, Zügeln, Zweigen, cippus, Asklepiosstatue, Ellenbogen, Lorbeerzweigen, Ruderern, Antefixen, Säulenbasis, SchuppenPanzer, Perlenquadrat, kerykeion, Adlerkopf, Türme, Erzen, Säulenkapitell, Göttinnenkopf, Panther-Biga, Asklepios-Statuette	Hermes, schreitender
PLANT	Olivenbaums, Kore, Früchten, Mohnblume, Bäumen, Weintrauben, Blättern	
ANIMAL	Hirsches, Ziegenbockes, Kretischen Stier, Nemeischen Löwen, Hirschen, Ebers, Krabbenscheren, Greifes, Asklepios-Schlange, Pferden, Pegasos-, Stieren, Ziegenbocks, Hahns, Stieres, Erymanthischen Eber	Elefantenzahn, darüberAdler, ihrPferd
VERBS	ringeln, gelagerte, getragen, umwundenen, spielende, fliegende, bekränzte, sitzende, gespanntem, springendem, liegende, liegenden, schreitendem, bekränzter, ringelnde, schreitende, gedrehten, geringelt, knieend, ausgestreckten, stehendem, emporringelt, geschultertem, gespannten, stützend, abgestützt, ausgestreckter, bekränztem, gestreckt, lagern, hängendem, umwundene, stehnend, umwundener, emporringelnde, drückt, gelagerter, thront, laufenden, kniet, schreitenden, gestützten, springenden, zieht, befreien, geringelten, schwimmendes, liegender, springende, gehaltenen, erhebt, thronenden, liegendem, stehendes, hängen, gesetzt, gespannt, windender, liegt, lehnt, stehenden, gelehnte, windenden, bekränzten, stehende, gelaufen, setzt, ringelnden, hängenden, gelehnt, Liegender, sitzender, fahrenden, gelagerten, thronender, stehender, abstützend	

A.3 Übersicht der nur von der manuellen Annotation gefundenen Entitäten (Deutsch)

Tabelle 11: Übersicht der nur von der manuellen Annotation gefundenen Entitäten (Deutsch)

Kategorie	Korrekt annotiert	Fehlerhafte annotiert
PERSON	Alexander des Großen , Kotys IV, Faustina I, Gordian III, Stadtgöttinnen, Hades	
OBJECT	Cista, Standarten	
PLANT		
ANIMAL		
VERBS		

A.4 Übersicht der nur von der manuellen Annotation gefundenen Entitäten (Englisch)

Tabelle 12: Übersicht der nur von der manuellen Annotation gefundenen Entitäten (Englisch)

Kategorie	Korrekt annotiert	Fehlerhafte annotiert
PERSON	Seleucus I	
OBJECT	branches, Cista, Herm, cista	
PLANT		
ANIMAL		
VERBS	leaned, Sailing	

A.5 Vergleich der alternativen Lemmatisierungs- und Stemmingalgorithmen der englischen Annotation

Tabelle 13: Übersicht der gefundenen Entitäten von Wordnet + Snowball und der *Kombi-Annotation*

	Nur von Wordnet + Snowball gefunden	**Nur von Kombi-Annotation gefunden**
Korrekte Annotation	OBJECT: pilei, crossing	ANIMAL: Beehive VERBS : rests, held, drawn
Fehlerhafte Annotation	OBJECT: sailing, crossing	

Tabelle 14: Übersicht der gefundenen Entitäten von SpaCy + Krovetz und der *Kombi-Annotation*

	Nur von SpaCy + Krovetz gefunden	**Nur von Kombi-Annotation gefunden**
Korrekte Annotation	OBJECT: Branches, Herm VERBS: Sailing	OBJECT: Helmeted, clothing, cornucopiae, bases, clothes, Beehive PERSON: Ganymede, figur, Seleucus I Nikator, Senate PLANT: branches VERBS: advanced
Fehlerhafte Annotation	OBJECT: crossing PERSON: dancing, African, Macedonian, Dancing, Seleucus I	OBJECT: Sailing, Hermes PERSON: herm, Herm

Tabelle 15: Übersicht der gefundenen Entitäten von Wordnet + Krovetz und der *Kombi-Annotation*

	Nur von Wordnet + Krovetz gefunden	**Nur von Kombi-Annotation gefunden**
Korrekte Annotation	OBJECT: pilei , Herm VERBS: Sailing	OBJECT: Helmeted, clothing, bases, cornucopiae, clothes, Beehive, Veiled PERSON: Ganymede, figur, Seleucus I Nikator, Senate VERBS: stands, held, hold, rests, stand, leaned on, advanced, fly, covered, set, holds, receive, drawn, curled
Fehlerhafte Annotation	Plant: leaves PERSON: dancing, African, Macedonian, Dancing, Seleucus I	OBJECT: Sailing, Hermes PERSON: herm, Herm VERBS: croddes

A.6 Vergleich der alternativen Lemmatisierungs- und Stemmingalgorithmen der deutschen Annotation

Tabelle 16: Übersicht der gefundenen Entitäten von HanTa + Cistem und der *Kombi-Annotation*

	Nur von HanTa + Cistem gefunden	**Nur von Kombi-Annotation gefunden**
Korrekte Annotation	OBJECT: Standarten VERBS: vorgestreckten, ausstreckend, vorgestreckt	OBJECT: Basen PLANT: Ast, Korn VERBS: abgestützt, erhoben, kniend
Fehlerhafte Annotation	OBJECT: Lorbeerkranzin, Felsten, Hera PERSON: M. , Faust, A. Korn, A	VERBS: Knien

Tabelle 17: Übersicht der gefundenen Entitäten von SpaCy + Pystem und der *Kombi-Annotation*

	Nur von SpaCy + Pystem gefunden	**Nur von Kombi-Annotation gefunden**
Korrekte Annotation	OBJECT: Akrotere, Akroteren, Standarten	OBJECT: Mitras, Rammsporen, Erzen, Knieen PERSON: Hero, Heroen, Apollon Smintheus VERBS: gelagerter, fasst, sitzenden, windender, laufenden, windenden, fliegende, emporringelt, gelehnte, gelagerten, schreitende, sitzende, bekränzter, gestützten, knieend, stehnend, gespannten, stehender, sitzender, bekränzt, stehenden, stehende, schwimmendes, abstützend, gedrehten, gehaltenen, stehendes, schwimmender, hängendem, gelagerte, schreitet, hängenden, springendem, gespanntem, geringelten, schreitenden, geschultertem, bekränztem, stehendem, springende, bekränzte, bekränzten, springenden
Fehlerhafte Annotation		OBJECT: schreitender PLANT: Kore PERSON: Fängen, Tanzende

Tabelle 18: Übersicht der gefundenen Entitäten von HanTa + Pystem und der *Kombi-Annotation*

	Nur von HanTa + Pystem gefunden	**Nur von Kombi-Annotation gefunden**
Korrekte Annotation	OBJECT: Standarten, Akrotere, Akroteren VERBS: vorgestreckten, ausstreckend, vorgestreckt	OBJECT: Basen, Mitras, Ast, Rammsporen VERBS:stützend, thront, laufenden, drückt, gelagerten, springend, gespannten, stehendem, laufend, setzt, bekränztem, fütternd, gesetzt, abstützend, springende, gedrehten, erhebt, getragen, gespannt, fliegend, geringelt, ringelt, gelagerter, sitzenden, schreitende, ringeln, sitzend, stehnend, knieend, bekränzte, bekränzten, kniet, springenden, gelagert, geringelten, schreitenden, schreitend, haltend, emporringelt, gelehnte, befreiend, stehend, bekränzter, sitzender, bekränzt, stehenden, stehendes, zieht, abgestützt, gelagerte, kniend, gespanntem, gelehnt, windender, fliegende, sitzende, lehnt, lagern, stehender, packend, stehende, schwimmender, brechend, windend, schwimmend, springendem, hängenden, schwimmendes, erhoben, gehaltenen, drehend, windenden, gestützten, hängendem, geschultertem, gestemmt
Fehlerhafte Annotation	OBJECT: Lorbeerkranzin, Felsten, Hera PERSON: M. , Faust, A., Korn, A	PLANT: Kore OBJECT: schreitender PERSON: Tanzende, Fängen VERBS: Knien

B Abbildungsverzeichnis

Abbildungsverzeichnis

1 Entnommen aus Huddleston und Pullum, 2002 - Konjugierte Verbformen in Englisch 9
2 Selbsterstellte Grafik - Ablauf des Programmes 15
3 Selbsterstellte Grafik - Annotation des Beispielsatzes 18
4 Selbsterstellte Grafik - Vergleich der Ergebnisse durch Lemmatisierung und Stemming (Deutsch) 22
5 Selbsterstellte Grafik - Ergebnis der Berechnung des Recallwertes zwischen den Modellen und den manuellen Annotationen der deutschen Daten 22
6 Selbsterstellte Grafik - Ergebnis der Berechnung des Precisionwertes zwischen den Modellen und den manuellen Annotationen der deutschen Daten 23
7 Selbsterstellte Grafik - Vergleich der Ergebnisse durch Lemmatisierung und Stemming (Englisch) 24
8 Selbsterstellte Grafik - Ergebnis der Berechnung des Recallwertes zwischen den Modellen und den manuellen Annotationen der englischen Daten 25
9 Selbsterstellte Grafik - Ergebnis der Berechnung des Precisionwertes zwischen den Modellen und den manuellen Annotationen der englischen Daten 25
10 Selbsterstellte Grafik - Vergleich des Recalls der Ergebnisse der NLP-Methoden (Deutsch) 30
11 Selbsterstellte Grafik - Vergleich der Precision der Ergebnisse der NLP-Methoden (Deutsch) 30
12 Selbsterstellte Grafik - Vergleich des Recalls der Ergebnisse der NLP-Methoden (Englisch) 31
13 Selbsterstellte Grafik - Vergleich der Precision der Ergebnisse der NLP-Methoden (Englisch) 32

C Literatur

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python* (1. ed.). O'Reilly.

Caumanns, J. (1997). *A Fast and Simple Stemming Algorithm for German Words* (Techn. Ber.). Freie Universität Berlin.

Davies, G. (2002). *A History of Money* (3. ed.). University of Wales Press.

Deligio, C., & Gencer, K. (2021). *Natural Language Processing auf mehrsprachigen Münzdatensätzen* (Master Thesis). Goethe Universität. Frankfurt am Main.

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press.

Imo, W. (2016). *Grammatik: Eine Einführung* (2016. Aufl.). J.B. Metzler Verlag.

Klinger, P. (2018). *Natural Language Processing to enable semantic search on numismatic descriptions* (Bachelor Thesis). Goethe Universität. Frankfurt am Main.

Krovetz, R. (1993). Viewing morphology as an inference process. *Proceedings of the 16th annual international ACM SIGIR conference on Researchand development in information retrieval*, 191–202.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press.

Michel, S. (2020). *Morphologie*. Narr Francke Attempto Verlag GmbH + Co. KG.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems*, *14*(3), 130–137.

Wartena, C. (2019). A Probabilistic Morphology Model for German Lemmatization. *Proceedings of the 15th Conference on Natural Language Processing*, 40–49.

Weissweiler, L., & Fraser, A. (2018). Developing a Stemmer for German Based on a Comparative Analysis of Publicly Available Stemmers. In G. Rehm & T. Declerck (Hrsg.), *Language Technologies for the Challenges of the Digital Age* (S. 81–94). Springer International Publishing.