|  |
| --- |
| DeNpKongru – Ein Erkenner zur Bestimmung der nominalen Kongruenz in deutschen Nominalphrasen aus Lernerkorpora |
| **Christopher Michael Chandler** |
| [christopher.chandler@rub.de](mailto:christopher.chandler@rub.de) |

Abstract

Nominale Kongruenz im Deutschen bereitet Lernern der deutschen Sprache etliche Schwierigkeiten aufgrund der Komplexität des deutschen Kasussystems. Aufgrund dessen wurden einfache heuristische Methoden entwickelt, um die Kongruenz innerhalb der Nominalphrasen zu bestimmen. Um diese Problematik besser darstellen zu können, wurden Anglophonen und Frankophonen in Bezug auf ihre Deutschkenntnisse miteinander verglichen. Es wird angenommen, dass Frankophonen weniger Probleme mit Nominalphrasen als Anglophonen hätten, da das Kongruenzprinzip im Französischen stärker vertreten ist als im Englischen. Um dieser Annahme auf den Grund zu gehen, wurden entsprechende Nominalphrasen aus dem Lernerkorpus Merlin extrahiert und analysiert. Die hier erzielten Ergebnisse bezüglich der Eingangshypothese sind nicht eindeutig und somit lässt sich nicht sagen, ob die Hypothese anzunehmen bzw. zu verwerfen ist, wobei die Ergebnisse leicht andeuten, dass beide Sprechergruppen fast gleich stark sind, was Kongruenz im Deutschen betrifft. Man kann jedoch anhand der Ergebnisse sehen, dass um so ein Projekt durchzuführen, könnte es von Vorteil sein, ein Korpus zu analysieren, das für so eine Aufgabestellung angefertigt wurde, denn so könnte man aussagekräftigere Ergebnisse erzielen.

1. Einleitung

Kongruenz als eine grammatikalische Notwendigkeit im Sinne von Satzkorrektheit ist in vielen Sprachen vertreten. Das gilt beispielsweise bei europäischen Sprachen (Englisch, Deutsch, Isländisch , Polnisch, Russisch, Spanisch), semitischen Sprachen (Arabisch, Hebräisch), uto-akztekischen Sprachen (Hopi) und auch bei isolierten Sprachen (Baskisch) (Corbett, 2003; Hofherr, 2020). Man kann also daraus schließen, dass die Kongruenz an sich ein sprachübergreifendes Phänomen ist.

Kongruenz im Allgemeinen bezeichnet das regelhafte linguistische Phänomen, dass bestimmte linguistische Elemente, wie z.B. Verben, Nomen, etc. mit anderen linguistischen Elementen, wie z.B. Pronomen, Adjektiven, etc. innerhalb einer linguistischen Äußerung miteinander übereinstimmen müssen (Corbett, 2003; Hofherr, 2020). Um das näher definieren zu können, kann man also in diesem Fall von morphologischer Kongruenz sprechen.

Da Kongruenz ein so stark präsentes Phänomen ist, ist es von Bedeutung festzulegen, wie man sie am besten beschreiben und untersuchen kann, denn Kongruenz bereitet nicht nur Nicht-Muttersprachlern große Schwierigkeiten, sondern auch Sprachsystemen, wo z.B. die Semantik-, oder Syntaxerkennung oft von Kongruenz abhängig ist (Corbett, 2003; Spinner & Juffs, 2008).

1. Kongruenz

Kongruenz kann verschiedene Ausprägungen annehmen und wie diese Ausprägungen ausfallen hängt direkt von der jeweiligen Sprache ab. Deswegen lässt sich nicht verallgemeinert sagen, wie genau die Kongruenz auszusehen hat und welche Formen es geben soll, dennoch gibt es ein paar Kongruenzarten, die hier aufgeführt werden, um diese zu veranschaulichen.

Um die Kongruenz und deren Merkmale klarer zu untersuchen wird als Objektsprache das Deutsche herangezogen, da diese Sprache auch im Laufe des Projekts als Objektsprache dient.

Bei beispielsweise Subjekt-Verb-Kongruenz müssen Subjekt und Verb bezüglich Numerus und Kasus miteinander übereinstimmen (siehe Tabelle 1) (Hofherr, 2020).

|  |  |
| --- | --- |
| Kongruenz | Nicht-Kongruenz |
| Ich gehe | Ich gehst\* |
| 1 SG.NOM gehen.pres1 SG | 2 SG.NOM  gehen.pres1 SG |

1. Beispiel der Subjet-Verb-Kongruenz

*Anmerkung*. In Anlehnung an *Agreement Morphology*, von P. C. Hofherr, 2020, S. 2. Copyright durch Oxford Research Encyclopedia of Linguistics.

Diese Form von Kongruenz (siehe Tabelle 1) ist in einigen Sprachen (Französisch, Spanisch, Russisch) vorhanden (Hofherr, 2020). Somit dürfte sie auch den meisten Sprechern der erwähnten Sprachen vertraut vorkommen.

Eine andere Variante der Kongruenz (Siehe Tabelle 2) ist die Kongruenz bei Nomen, wo Numerus, Genus und Kasus alle eine Rolle in einer linguistischen Phrase spielen. Diese wird nachfolgend als nominale Kongruenz bezeichnet. Bei der nominalen Kongruenz spielt der Kopf der Nominalphrase eine entscheidende Rolle, denn alle betroffenen linguistischen Elemente müssen mit dem Kopf kongruieren (Hofherr, 2020).

|  |  |
| --- | --- |
| Kongruenz | Nicht-Kongruenz |
| Der kleine Hund | Die kleine Hund\* |
| 1 SG.NOM bestimmt.mask.sing Artikel + mask.sing.nomen + mask.sing.nominativ adjektiv | 1 SG.NOM bestimmt.fem.sing Artikel + fem.sing.nomen + fem.sing.nominativ adjektiv |

1. Beispiel der nominalen Kongruenz

*Anmerkung*. In Anlehnung an *Agreement Morphology*, von P. C. Hofherr, 2020, S. 2. Copyright durch Oxford Research Encyclopedia of Linguistics.

Wenn ein Element in einer Nominalphrase fehlt oder nicht richtig eingesetzt wurde, ist die gesamte Äußerung falsch. Im Vergleich zu anderen Sprachen wie z.B. Englisch, Italienisch, Französisch, da so ein entsprechendes Kasussystem fehlt, gilt die deutsche Sprache als bei Nicht-Muttersprachlern bzw. Lernenden als schwierig (Spinner & Juffs, 2008).

Um die genauen Schwierigkeiten in Bezug auf nominale Kongruenz im Deutschen bestimmen zu können, werden Texte von Anglophonen und Frankophonen aus dem Lernerkorpus Merlin (Wisniewski et al., 2018) extrahiert. Gewählt wurden diese beiden Gruppen unter der Annahme, dass Frankophonen am Anfang eher weniger Schwierigkeiten bei nominaler Kongruenz als Anglophonen haben, da Kongruenz, auch wenn weniger komplex, im Französischen stärker präsent ist als im Englischen. Es wird jedoch erwartet, dass diese Differenz mit steigendem Niveau abnimmt.

Die Ergebnisse sollen Auskunft darüber geben, inwieweit man Strategien oder Ansätze entwickeln kann, um Lernenden effektivere Lernmethoden bezüglich des Erwerbs der nominalen Kongruenz im Deutschen anbieten zu können.

1. Verwandte Literatur

Die Kernthematik bezieht sich hier auf morphologische Kongruenz, insbesondere Kongruenz bei Nominalphrasen bzw. nominale Kongruenz. In folgenden Kapiteln wird ein regelbasiertes Vorgehen aufgestellt, jedoch gibt es schon LSTM-basierte Ansätze (Ravfogel et al., 2018), um Kongruenz bei Phrasen zu bestimmen, wo nominale Kongruenz und (Subjekt)-Verb-Kongruenz eine Rolle spielen. Das RNN-Model benutzt lokale Heuristiken, die aus einer baskischen Phrase extrahiert werden können, um die Kongruenz zu bestimmen. Nichtsdestotrotz sind fundierte grammatikalische Kenntnisse der Sprache unerlässlich, wenn die Kongruenz aufgeschlüsselt werden sollte (Ravfogel et al., 2018).

Bei Sprachen, wo die morphologischen Merkmale wesentlich weniger Komplexität ausweisen, wie z.B. Englisch, wurde auch Kongruenzbestimmung an einem LSTM-Ansatz erprobt. LSTMs weisen zwar eine sehr hohe Accuracy bzw. Genauigkeit auf, dennoch gilt dies nur bei Äußerungen einer gewissen Länge, bei der strukturelle und sequentielle Informationen nicht miteinander harmonieren (Linzen et al., 2016).

1. Nominale Kongruenz

Die Kongruenzart, die im Rahmen dieses Projekts untersucht werden soll, ist die nominale Kongruenz. Ziel ist es, automatisch bestimmen zu können, ob eine Nominalphrase korrekt bzw. kongruent oder inkorrekt bzw. nicht-kongruent ist.

Eine Nominalphrase gilt als korrekt, wenn alle ihre Elemente mit dem Kopf der Nominalphrase kongruieren bzw. übereinstimmen. Zudem liegt eine Kongruenz nicht vor, wenn mindestens ein Element nicht mit dem Kopf kongruiert.

* 1. Termini

Um die Kongruenz genauer untersuchen und beschreiben zu können, müssen bestimmte Termini festgelegt werden. Dazu kann man die Interaktion der einzelnen Elemente in Tabelle 3 einsehen.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Domäne | |  |  |  |  |
|  | Ziel |  | Kontroller |  |  |  |
|  | Das blaue |  | Haus |  |  | Bedingung |
|  |  |  |  |  |  |  |
|  | Merkmal:  Numerus, Genus, Kasus  Wert: Singular, Neutrum, Nominativ | | |  |  |  |

1. Umriss der Kongruenz-Termini

*In Anlehnung* an *Agreements: Terms and Boundaries,* von C. Greville, S. 2., 2003. Copyright durch SMG conference papers.

Der Kontroller bestimmt die Kongruenz und das Ziel muss mit den Eigenschaften des Kontrollers übereinstimmen. Die Domäne ist die syntaktische Umgebung, worin die Nominalphrase sich befindet. Merkmale sind die Eigenschaften der Kongruenz und es gibt Bedingungen außerhalb der Domäne, die Auswirkungen auf die Kongruenz haben (Corbett, 2003; Hofherr, 2020).

* 1. Intention des Sprechers

Obwohl Zielhypothesen der einzelner Texte gemacht wurden, in dem eine korrigierte Version festgelegt wurde (Boyd et al., 2014), wird die Intention des Sprechers bei der Bestimmung der Nominalphrasenarten nicht berücksichtigt. Der Hauptgrund dafür ist, dass diese sich nicht so leicht automatisch bestimmen lässt und mehrere Möglichkeiten als richtig gelten könnten:

|  |  |
| --- | --- |
| Satzarten | Sätze |
| Lernersatz | Ich sehe die Hund\* |
| Zielhypothese 1 | Ich sehe den Hund |
| Zielhypothese 2 | Ich sehe die Hunde |

1. Zielhypothesen

Wenn man beispielsweise (siehe Tabelle 4) annimmt, dass die Nominalphrase Ich sehe die Hund\* falsch ist, hat man zwei Möglichkeiten, den Satz zu korrigieren. Ohne Weiteres ist die Intention des Lerners unbekannt und somit wird sie hier nicht berücksichtigt. Es ist daher möglich, die Annahme zu vertreten, dass man den Satz als solchen behandeln soll.

1. Vorgehensweise

In diesem Projekt geht es hauptsächlich darum, die Kongruenz der einzelnen Nominalphrasen zu bestimmen, die aus bestimmten Lernerkorpustexten extrahiert wurden. Nominalphrasen bestimmen zu können setzt voraus, dass der Eingangstext schon entsprechend tokenisiert und nach Wortklasse getaggt wurde. Diese Aufgabe wurde im Vorfeld von dem ParZu-Tagger (Sennrich et al., 2009)

übernommen, sodass die Texte im Rahmen des

Projekts nicht geparst werden mussten.

Folgende Wortklassen nach dem STTS-Schema (Schiller et al., 1999) sind für die Nominalphrasen relevant: NOUN, PREP, ART und ADJA. Ausgeschlossen wurden alle anderen Wortklassen.

Nach dem erfolgreichen Parsen wurden die daraus stammenden Conll-Dateien nach geeigneten Nominalphrasen untersucht. Geeignete bedeutet an dieser Stelle, dass mindestens ein Nomen in einer Äußerung unbestimmter Länge vorkommen muss.

Die Nominalphrasen wurden extrahiert und in einer gesonderten Datei gespeichert. Ein Nominalphrase-Eintrag hat dann beispielsweise die folgende Form:

|  |  |
| --- | --- |
| Nominalphrase | meine Fahrkarte |
| Morphologische Information | meine ART Fem|Acc|Sg,  Fahrkarte N Fem|Acc|Sg |
| Satz | Ich gebe Dir meine Fahrkarte . |

1. Nominalphraseneintrag
   1. Kongruenz kodieren

Damit die Kongruenz festgehalten und später nachvollzogen werden kann, wurde die Kongruenz-Art kodiert. Nominalphrasen können verschiedene Formen annehmen und somit unterschiedlichen Kategorien zugeordnet werden. Die Nominalphrasen wurden erst in unterschiedliche Kategorien eingeteilt. Innerhalb dieser Kategorien wurden einfache Heuristiken angewandt, um die Kongruenz bestimmen zu können. Die Kategorien kann man Tabelle 6 entnehmen.

Nachdem eine Nominalphrase einer Kategorie zugeordnet wird, wird sie innerhalb dieses Kontexts auf Kongruenz untersucht. Dazu wurden Kongruenz-codes vergeben, sodass festgelegt werden kann, ob Kongruenz vorhanden ist und wenn ja , um welche Art es sich handelt (Siehe Tabelle 6, 7 und 8). Wenn die Kongruenz innerhalb einer Kategorie nicht gilt, wird die Nominalphrase mit einem Code versehen, der angibt, dass die Kongruenz nicht zutrifft (siehe Tabelle 8).

|  |  |
| --- | --- |
| Arten | Beispiel |
| EINFACH | Stadt |
| ART | Das Leben /  Welche Freizeitaktivitäten, |
| PREP | Mit Kindern  Für neuen Aufgaben\*,  In der Woche |
| Rechtschreibfehler | Reche Pfliche\*,  Wiviel\* |
| Eigennamen | Katharina,  Maria Meier |
| Redewendungen | Liebe Julia,  Mit freundlichen Grüßen |

1. Kongruenzarten

|  |  |
| --- | --- |
| Code | Bedeutung |
| 0 | EINFACH |
| 1 | ART |
| 2 | PREP |
| 3 | Eigennamen |
| 4 | Redewendung |

1. Kongruenz-Codes

|  |  |
| --- | --- |
| Code | Bedeutung |
| 10 | EINFACH |
| 11 | ART |
| 12 | PREP |
| 99 | Unbekannt |

1. Nicht-Kongruenz-Codes
   1. Beispiel der Kongruenz

Es werden einfache Heuristiken angewandt um festzustellen, ob eine Nominalphrase kongruiert. Die einfachste Nominalphrase zu bestimmen ist EINFACH. Kommt eine Nominalphrase ohne Artikel bzw. Determiner vor und wird sie richtig geschrieben, so wird sie als richtig eingestuft. Wenn das nicht der Fall ist, kann man entweder von einem Neologismen oder einem Rechtschreibfehler ausgehen.

ART ist eine komplexere Nominalphrase, die einen Determiner oder Possiv-Artikel enthält. Kongruieren Determiner und Nominalphrase, dann soll sie als richtig gelten, ansonsten als falsch.

PREP folgt einem ähnlichen Ansatz wie bei ART, aber die Präpositionen lösen eine bestimmte Kongruenz aus, z.B. Mit verlangt Dativ. Wenn alle in Kasus übereinstimmen, was die Präpositionen bestimmt, dann gilt die Nominalphrase als richtig. Wenn nicht, dann ist sie falsch.

Eigennamen und Redewendungen werden auf Rechtschreibung untersucht. Wenn die Rechtschreibung stimmt, werden sie entsprechend als EIGENNAMEN oder REDEWENDUNGEN markiert.

|  |  |
| --- | --- |
| Code | Bedeutung |
| 0 | Eine Nominalphrase, die ohne Determinier vorkommt. |
| 1 | Eine Nominalphrase, die mit einem Determinier und evtl. Adjektiven vorkommt. |
| 2 | Eine Nominalphrase mit Präpositionen und evtl. Adjektiven |
| 3 | Eigennamen; Namen von Personen, Städten, Firmen, etc. |
| 4 | Redewendungen oder Anreden, die in der deutschen Sprache üblich sind |
| 10 | Die Nominalphrase wird der Kategorie 0 zwar zugeordnet, aber die Kongruenz ist falsch.  Hier werden Rechtschreibfehler festgelegt. |
| 11 | Die Nominalphrase wurde der Kategorie 1 zwar zugeordnet, aber die Kongruenz ist falsch.  Hier werden Kongruenzfehler festgelegt, z.B. Der Haus\*. |
| 12 | Die Nominalphrase wurde der Kategorie 2 zwar zugeordnet, aber die Kongruenz ist falsch.  Hier werden Kongruenzfehler festgelegt, die aufgrund der Präpositionen nicht stimmen z.B. Mit Der Haus. |
| 99 | Eine Kongruenzart konnte nicht bestimmt werden. |

1. Kongruenzerklärungen
2. Datensatz

Der Merlin-Korpus wird als Datensatz für das Projekt verwendet. Es ist ein fehlerannotiertes schriftliches Lernerkorpus für Deutsch, Italienisch und Tschechisch (Wisniewski et al., 2018). Lerner unterschiedlicher Niveaus wurden nach dem gemeinsamen europäischen Referenzrahmen für Sprachen bzw. CEFR bewertet. Alle Texte wurden nach einer vorgegebenen Aufgabe z.B. Alltagsszenarien oder Arbeitsszenarien verfasst (Wisniewski et al., 2018). Diese wurden dann anschließend bewertet und ggf. korrigiert.

Die Texte, die in italienischer und tschechische Sprache verfasst wurden, werden in diesem Projekt nicht berücksichtigt. Deswegen werden nur die deutschen Texte analysiert, die von Lernern geschrieben wurden.

|  |  |  |
| --- | --- | --- |
| Fair CEFR Level | | Deutsch |
| Anfänger | A1 | 57 |
| A2 | 199 |
| A2+ | 107 |
| Selbständig | B1 | 219 |
| B1+ | 115 |
| B2 | 219 |
| B2+ | 73 |
| Fortgeschritten | C1 | 42 |
| C2 | 4 |
| Textanzahl | | 1,035 |
| Wortanzahl | | 125,927 |

1. Merlin-Korpus-Textanzahl

Da nur 4 Texte auf C2 Niveau verfasst wurden, werden diese Texte auch nicht berücksichtigt. Es bleiben also Texte von A1 bis C1 zum Analysieren.

* 1. Aufbau der einzelnen Einträge

Um die Texte nachher auswerten und sortieren zu können, wurde jeder Text mit Metadaten versehen (siehe Tabelle 11 u. Tabelle 12). Zusätzlich zum Lernertext wurde für einige Texte in diesem Datensatz eine korrigierte Version bzw. eine Zielhypothese angelegt (Boyd et al., 2014), jedoch wird diese bei der Analyse im Rahmen dieses Projekts nicht berücksichtigt (Siehe Kapitel 4.2).

* 1. Projektdatensatz

Zu den Muttersprachen der Teilnehmer gehören die folgenden Sprachen: Arabisch, Chinesisch, Tschechisch, Englisch, Französisch, Ungarisch, Italienisch, Polnisch, Portugiesisch, Russisch, Spanisch und Türkisch. Einige Sprecher haben entweder ihre Muttersprache nicht angegeben oder sie wurde nicht erfasst. Relevant für dieses Projekt

sind die Texte, die von Anglophonen und Frankophonen verfasst wurden (siehe Tabelle 11).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | B1 | B2 | C1 | Gesamt |
| English | 15 | 1 | 11 | 9 | 10 | 46 |
| French | 5 | 1 | 2 | 19 | 12 | 39 |
| Gesamt | 20 | 2 | 13 | 28 | 22 | 85 |

1. Texte von Anglophonen und Frankophonen
2. Evaluation

Da die Daten bzw. Texte für die jeweiligen Gruppen nicht gleichmäßig erhoben wurden, werden die Datensätze für die Trainings- und Testphase einigermaßen gleich aufgeteilt und somit wird nur ein Bruchteil der Daten benutzt.

Es handelt sich hier um ein regelbasiertes Model und nicht um ein statistisches oder RNN-Model, weshalb das Programm an sich nicht trainiert werden muss. Jedoch werden die Textdaten so aufgeteilt, dass ungefähr 50% als Trainingsdaten dienen, um die Algorithmen zu verbessern und die restlichen 50% als tatsächliche Testdaten.

* 1. Trainingsphase

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | B1 | B2 | C1 | Gesamt |
| English | 2 | 0 | 1 | 4 | 5 | 12 |
| French | 2 | 0 | 1 | 4 | 5 | 12 |
| Gesamt | 4 | 0 | 2 | 8 | 10 | 24 |

1. Trainingsdaten

Während der Trainingsphase wurde die Performanz des Programms anhand Scikit learn (Buitinck et al., 2013) bemessen. Es wurden 6 Dateien als Referenz-Dateien verwendet, woraus die Gold-Dateien manuell erstellt wurden. In Tabelle 15 sieht man, dass das Programm 50% Genauigkeit bei 201 Datenpunkten aufweist. Was auffällt ist, dass falsche einfache nominale Phrasen gar nicht erfasst werden. Die Gründe hierfür sind zweierlei. Erstens kommen solche Nominalphrasen sehr selten vor und werden somit nur äußerst selten erfasst. Zweitens wurden am Anfang einige Nominalphrasen automatisch aussortiert, weil nur Nominalphrasen vom Algorithmus berücksichtigt wurden, wenn sie mit den Sätzen ausgegeben wurden. Das heißt, sie mussten eine Struktur aufweisen, die in Tabelle 5 zu sehen ist. Da die Nominalphrasen somit fehlerhaft waren, konnten sie nicht analysiert werden.

Ein andere Auffälligkeit ist, dass sehr viele Nominalphrasen als Unbekannt markiert wurden. Das hängt vor allem damit zusammen, dass der ParZu-Tagger (Sennrich et al., 2009) die Nominalphrasen nicht richtig bzw. vollständig erfassen konnte. Es entstanden also Nominalphrasen wie z.B. *Grüßen Anlage* oder *Erfahrung in nach einer* (Siehe Tabelle 16).

Andere Schwierigkeiten, die beim Training aufgetaucht sind, war die Frage, wie Adressen wie z.B. *Schmidt Tobias Müllergasse* behandelt werden sollen. Solche Ausdrücke sind per se immer richtig und aufgrund dessen im Rahmen dieses Projekts uninteressant. Bei vielen Texten war das Ziel, jemanden anzuschreiben oder auf eine Stellenanzeige zu antworten (Boyd et al., 2014). Das führt dann dazu, dass man sehr viele Adressen verwendet. Diese werden auch vom Programm als Nominalphrasen erkannt und erfasst (siehe Tabelle 17), dennoch sind sie ebenfalls für das Projekt uninteressant.

Zu der Formalität des Textes gehört auch, dass man Redewendungen verwendet (Siehe Tabelle 18). Redewendungen bzw. Anreden sind komplexer als Adressen, denn sie müssen korrekt wiedergegeben werden. Daran kann man sich auch orientieren, wenn man die Kongruenz bestimmen möchte. Sätze wie z.B. Mit freundlichen Grüßen, Lieber Max, etc. haben zwar zum Teil eine festgelegte Form, dennoch muss man auf die Kongruenz achten, wenn die Äußerung als solche auch als korrekt gelten soll.

Rechtschreibfehler sind zu erwarten (siehe Tabelle 6) und haben Auswirkungen auf die Ergebnisse. Wie im Kapitel 4.2 und in Tabelle 8 erwähnt, ist es nicht immer möglich, sich die richtige Intention des Sprechers zu erschließen. Dazu zählen zum Beispiel Neuschöpfungen wie z.B., der Informatik-Club. Um solche Vorkommnisse besser erkennen zu können, müssen robustere Algorithmen eingesetzt werden, die zwischen Neuschöpfungen und Rechtschreibfehler unterscheiden können. Des Weiteren muss es auch möglich sein, einen Rechtschreibfehler festzulegen, damit eine partielle Kongruenz (siehe Tabelle 8) umzusetzen ist.

In Hinblick auf die oben genannten Kenntnisse reduziert sich die möglichen Nominalphrasen auf eine kleinere Menge, dennoch wurden alle Gruppen berücksichtigt. Beim Analysieren der Durchschnittswerte der Trainingsphrase sieht man, dass sich die Zahl der unbekannten Nominalphrasen beim zunehmenden Sprachniveau steigert. Auch die Anzahl der falschen Nominalphrasen im Vergleich zu den richtigen Nominalphrasen ist überraschend, denn es wurden mehr falsche Nominalphrasen festgestellt als richtige.

* 1. Testphase

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A1 | A2 | B1 | B2 | C1 | Gesamt |
| English | 2 | 1 | 1 | 4 | 5 | 13 |
| French | 2 | 1 | 1 | 4 | 5 | 13 |
| Gesamt | 4 | 2 | 2 | 8 | 10 | 26 |

1. Testdaten

Der Datensatz, der während der Testphase verwendet wurde, war zwar größer, aber dafür balanciert. Leider war es während der Testphase auch nicht möglich, gleichmäßige Gruppen in Bezug auf die Sprachniveaus zu bilden.

Wenn man sich die Durchschnittswerte (Siehe Tabelle 20 – 29) ansieht, stellt man fest, dass die Anzahl der unbekannten Nominalphrasen beim steigenden Niveau auch zunimmt. Auffällig ist auch, dass die Anzahl der richtigen Nominalphrasen bzw. falschen Nominalphrasen verhältnismäßig gleich ist.

1. Diskussion

Im Rahmen des Projekts wurden Texte, die sowohl von Frankophonen als auch von Anglophonen geschrieben wurden, miteinander verglichen. Es wurde entschieden, ein regel- bzw. heuristikbasiertes Verfahren zu verwenden, um die sprachlichen Unterschiede zwischen den zwei Sprechergruppen aufzuschlüsseln. Die ursprüngliche Motivation dieses Projekts war die sprachlichen Unterschiede zwischen Anglophonen und Frankophonen in Hinblick auf Nominalphrasen beim Verfassen von Texten auf Deutsch zu untersuchen. Es wurde hypothetisiert, dass die Frankophonen am Anfang - sprich A1, A2 - weniger Fehler machen würden als die Anglophonen. Diese Unterschiede würden allerdings beim steigenden Niveau abnehmen.

Die Regeln, die aufgestellt wurden, um die Nominalphrasen in Kategorien einzuteilen, ist vom Prinzip her meiner Meinung nach zwar richtig, aber die Datengrundlage und die Parser-Methoden sind nicht dafür geeignet, solche Experimente durchzuführen. Im Allgemeinen lässt sich schwer sagen, ob so eine Methode geeignet ist, um Differenzen zwischen zwei Sprechergruppen festzulegen.

Zum einen ist das Korpus dafür zu klein und zum anderen waren die Nominalphrasen, die daraus entstanden sind, aus einigen Gründen zum Teil nicht aussagekräftig genug. Aufgrund dessen sind die Ergebnisse, die während dieses Projekts erzielt wurden, anders zu interpretieren.

Die Ergebnisse zeigen, dass man die Nominalphrasen anders aufbereiten oder ggf. mit einem kleineren Korpus hätte arbeiten sollen. Der Grund dafür ist, dass die Nominalphrasen an sich entweder zu klein oder nicht zahlreich genug waren, was am Ende dazu geführt hat, dass die Nominalphrasen, die gezählt wurden, in Bezug auf Aussagekraft der allgemeinen Ergebnisse nicht so stark ins Gewicht fallen konnten.

Viel auschlaggebender war das Framework, das im Laufe dieses Projekts entwickelt wurde. Das Framework bzw. die Pipeline eignen sich gut dafür, so ein Experiment durchzuführen. Mit anderen Methoden und einem anderen Korpus können bessere Ergebnisse erzielt werden.

Die Methoden sollten eventuell robuster und statistisch begründet sein. Das Korpus sollte aus Phrasen bestehen, die viele Nominalphrasen haben. Man könnte die Lerner dazu bringen, solche Phrasen zu schreiben, indem man ihnen eine andere Aufgabenstellung gibt, wie z.B. möglichst lange und detaillierte Sätze zu bilden, denn solche Sätze würden höchstwahrscheinlich auch viele Nomen und Adjektive enthalten.

1. Fazit und zukünftige Ansätze

Solche Experimente durchzuführen ist auf jeden Fall hilfreich, denn Kongruenz bzw. die morphologische Kongruenz ist nicht nur ein wichtiger Bestandteil der deutschen Sprache, sondern vieler europäischen Sprachen. Ergebnisse diesbezüglich tragen dazu bei, bessere Systeme für Lerner und Muttersprachler zugleich zu entwickeln. In zukünftigen oder ähnlichen Projekten soll der Fokus darauf liegen, entweder ein für so eine Aufgabe bestimmtes Korpus anzufertigen oder andere Parsing- und Kongruenzbestimmungsmethoden zu entwickeln.

Danksagung

Dieses Projekt wurde im Rahmen des computerlinguistischen Kurses *Korpuslinguistische Analysen der Nominalflexion im Deutschen (050041-SoSe23)*an der Ruhr-Universität Bochum in Deutschland entwickelt. Ideen und Code von den Mitkommilitonen und der Kursleiterin Frau Dr. Ronja Laarman-Quante wurden hierfür verwendet und waren bei der Planung und Ausführung des Projekts unentbehrlich.

Literaturverzeichnis

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B., & Vettori, C. (2014). The MERLIN corpus : Learner language and the CEFR. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, 1281‑1288. http://www.lrec-conf.org/proceedings/lrec2014/pdf/606\_Paper.pdf

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software : Experiences from the scikit-learn project. *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108‑122.

Corbett, G. (2003). Agreement : Terms and boundaries. *SMG conference papers*.

Hofherr, P. C. (2020). *Agreement Morphology*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.500

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521‑535. https://doi.org/10.1162/tacl\_a\_00115

Ravfogel, S., Goldberg, Y., & Tyers, F. (2018). Can LSTM Learn to Capture Agreement ? The Case of Basque. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 98‑107. https://doi.org/10.18653/v1/W18-5412

Schiller, A., Teufel, S., Stöckert, C., & Thielen, C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Universität Tübingen. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf

Sennrich, R., Schneider, G., Volk, M., & Warin, M. (2009). *A New Hybrid Dependency Parser for German*. https://doi.org/10.5167/uzh-25506

Spinner, P., & Juffs, A. (2008). L2 grammatical gender in a complex morphological system : The case of German. *Iral-international Review of Applied Linguistics in Language Teaching - IRAL-INT REV APPL LINGUIST*, *46*, 315‑348. https://doi.org/10.1515/IRAL.2008.014

Wisniewski, K., Abel, A., Vodičková, K., Plassmann, S., Meurers, D., Woldt, C., Schöne, K., Blaschitz, V., Lyding, V., Nicolas, L., Vettori, C., Pečený, P., Hana, J., Čurdová, V., Štindlová, B., Klein, G., Lauppe, L., Boyd, A., Bykh, S., & Krivanek, J. (2018). *MERLIN Corpus | Resources for research and practice related to foreign language learning* (1.1) [jeu de données]. https://www.merlin-platform.eu/

Eigenständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet habe. Ich erkläre weiterhin, dass ich alles gedanklich, inhaltlich oder wörtlich von anderen (z.B. aus Büchern, Zeitschriften, Zeitungen, Lexika, Internet usw.) Übernommene als solches kenntlich gemacht, d.h. die jeweilige Herkunft im Text oder in den Anmerkungen belegt habe. Dies gilt ggf. auch für Tabellen, Skizzen, Zeichnungen, bildliche Darstellungen usw.

Ich nehme zur Kenntnis, dass die nachgewiesene Unterlassung der Herkunftsangabe als versuchte Täuschung bzw. als Plagiat („geistiger Diebstahl“) gewertet wird. Unkenntnis der in der Wissenschaft gebräuchlichen Regeln gilt nicht als Entlastung.

Ich anerkenne hiermit, dass bei Vorliegen eines Plagiats die Arbeit nicht als selbständige Leistung gewertet wird mit der Folge, dass mein Anspruch auf einen Leistungsnachweis (auch Teilnahmenachweis) in der Lehrveranstaltung und die Möglichkeit einer Nachbesserung oder Wiederholung der Arbeit entfällt.

Christopher Michael Chandler

Ort, Datum

Anhang

|  |  |
| --- | --- |
|  | Definition |
| Author id | Identifikationsnummer des Texts |
| Test language | Die Sprache, worin der Text verfasst wurde |
| CEFR level of test | Sprachniveau der Aufgabe |
| Task | Die Aufgabe, nach der der Text verfasst wurde |
| Mother tongue | Muttersprache |
| Age | Alter |
| Gender | Geschlecht |
| Overall CEFR rating | Gesamtbewertung des Textes |
| Grammatical accuracy | Grammatikalische Richtigkeit |
| Orthography | Rechtschreibung in Bezug auf Tippfehler, Zeichensetzung, Groß- und Kleinschreibung, Wortgrenzen, etc. |
| Vocabulary range | Lexikalische wissen, lexikalische Fehler bzw. falsche Wortwahl |
| Vocabulary Control | Inwieweit man mit dem eigenen Wortschatz geben kann |
| Coherence/Cohesion | Konnektoren und Aufbau des Textes |
| Sociolinguistic  appropriateness | Anreden, Anfragen innerhalb eines gegebenes Kontexte |

1. Merlin-Text-Aufbau

|  |  |
| --- | --- |
|  | Definition |
| Author id | 1023\_0108888 |
| Test language | German |
| CEFR level of test | B2 |
| Task | Au pair writes letter of complaint to Agency - ID: 1023\_100303 |
| Mother tongue | French |
| Age | 27 |
| Gender | Female |
| Overall CEFR rating | B2 |
| Grammatical accuracy | B2 |
| Orthography | B2 |
| Vocabulary range | B2 |
| Vocabulary Control | B1 |
| Coherence/Cohesion | B2 |
| Sociolinguistic  appropriateness | B2 |

1. Merlin-Text-Aufbau - Beispiel

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| 0 - EINFACH | 0.91 | 0.50 | 0.65 | 20 |
| 01 - ART | 0.13 | 0.09 | 0.11 | 32 |
| 02 - PREP | 1.00 | 0.08 | 0.15 | 24 |
| 03 - Eigennamen | 0.38 | 0.38 | 0.38 | 26 |
| 04 - Redewendung bzw. Satz | 1.00 | 0.25 | 0.40 | 8 |
| 10 - EINFACH (Nicht Kongruenz) | 0.33 | 1.00 | 0.49 | 17 |
| 11 - ART (Nicht Kongruenz) | 0.00 | 0.00 | 0.00 | 9 |
| 12 - PREP (Nicht Kongruenz) | 0.75 | 0.35 | 0.48 | 17 |
| 99 - Unbekannt (Nicht Kongruenz) | 0.48 | 0.77 | 0.59 | 48 |
| Accuracy |  |  | 0.43 | 201 |
| Macro Avg | 0.55 | 0.38 | 0.36 | 201 |
| Weighted Avg | 0.53 | 0.43 | 0.39 | 201 |

1. DpNpKongru - Performanz

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Np-ID | Kongruenz-code | NP | Morphologische Information | Satz |
| 15\_54 | 99 | Mit freundlichen Grüßen Anlage | Mit,PREP,\_|\_|Dat|\_  freundlichen,ADJA,\_|\_|Dat|\_  Grüßen,N,\_|\_|Dat|\_  Anlage,N,\_|Fem|Dat|Sg | Mit freundlichen Grüßen , Anlage : Anmelungsschein in Kopie |
| 15\_55 | 99 | Grüßen Anlage | Grüßen,N,\_|\_|Dat|\_  Anlage,N,\_|Fem|Dat|Sg | Mit freundlichen Grüßen , Anlage : Anmelungsschein in Kopie |

1. 1023\_0108888 Unbekannte

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Np-ID | Kongruenz-code | NP | Morphologische Information | Satz |
| 1\_1 | 3 | Schmidt Tobias | Schmidt,N,\_|Fem|\_|Sg,Tobias,N,\_|Fem|\_|Sg | Schmidt Tobias Müllergasse….. |
| 1\_2 | 10 | Schmidt Tobias Müllergasse, | Schmidt,N,\_|Fem|\_|Sg,Tobias,N,\_|Fem|\_|Sg,Müllergasse,N,\_|Fem|\_|Sg | Schmidt Tobias Müllergasse….. |

1. 1023\_0108888 Adressen

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Np-ID | Kongruenz-code | NP | Morphologische Information | Satz |
| 1\_1 | 99 | Hallo Ingo | Hallo,N,\_|Neut|\_|Sg,Ingo,N,\_|Masc|\_|Sg | 16.03.2012 Hallo Ingo ! |
| 7\_17 | 99 | Mit Grußen Maria | Mit,PREP,\_|\_|Dat|\_,Grußen,N,\_|\_|Dat|Sg,Maria,N,\_|\_|Dat|Sg´ | Mit freundlich Grußen , Maria |

1. 1071\_0024802 Redewendungen

|  |  |  |  |
| --- | --- | --- | --- |
| A1\_EN\_AVG\_WAHR | A1\_EN\_AVG\_FALSCH | A1\_EN\_AVG\_UNBEKANNT | |
| 4 | 2 | 8 |  |
|  | training | averages | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| A1\_FR\_AVG\_WAHR | A1\_FR\_AVG\_FALSCH | A1\_FR\_AVG\_UNBEKANNT | |
| 5 | 4 | 6,5 |  |
|  | training | averages | 2 |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| A2\_EN\_AVG\_WAHR | A2\_EN\_AVG\_FALSCH | A2\_EN\_AVG\_UNBEKANNT | |
| UNK | UNK | UNK |  |
|  | training | averages | 3 |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| A2\_FR\_AVG\_WAHR | A2\_FR\_AVG\_FALSCH | A2\_FR\_AVG\_UNBEKANNT | |
| UNK | UNK | UNK |  |
|  | training | averages | 4 |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| B1\_EN\_AVG\_WAHR | B1\_EN\_AVG\_FALSCH | B1\_EN\_AVG\_UNBEKANNT | |
| 4 | 3 | 3 |  |
|  | training | averages | 5 |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| B1\_FR\_AVG\_WAHR | B1\_FR\_AVG\_FALSCH | B1\_FR\_AVG\_UNBEKANNT | |
| 3 | 5 | 10 |  |
|  | training | averages | 6 |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| B2\_EN\_AVG\_WAHR | B2\_EN\_AVG\_FALSCH | B2\_EN\_AVG\_UNBEKANNT | |
| 12 | 18,5 | 17,5 |  |
|  | training | averages | 7 |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| B2\_FR\_AVG\_WAHR | B2\_FR\_AVG\_FALSCH | B2\_FR\_AVG\_UNBEKANNT | |
| 9,5 | 17,75 | 23 |  |
|  | training | averages | 8 |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| C1\_EN\_AVG\_WAHR | C1\_EN\_AVG\_FALSCH | C1\_EN\_AVG\_UNBEKANNT | |
| 11,4 | 12,8 | 25,8 |  |
|  | training | averages | 9 |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| C1\_FR\_AVG\_WAHR | C1\_FR\_AVG\_FALSCH | C1\_FR\_AVG\_UNBEKANNT | |
| 9,6 | 8,6 | 26 |  |
|  | training | averages | 10 |

|  |  |  |  |
| --- | --- | --- | --- |
| A1\_EN\_AVG\_WAHR | A1\_EN\_AVG\_FALSCH | A1\_EN\_AVG\_UNBEKANNT |  |
| 1,5 | 3 | 4,5 |  |
|  | test | averages | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
| A1\_FR\_AVG\_WAHR | A1\_FR\_AVG\_FALSCH | A1\_FR\_AVG\_UNBEKANNT |  |
| 4,5 | 1,5 | 9 |  |
|  | test | averages | 2 |

|  |  |  |  |
| --- | --- | --- | --- |
| A2\_EN\_AVG\_WAHR | A2\_EN\_AVG\_FALSCH | A2\_EN\_AVG\_UNBEKANNT |  |
| 2 | 14 | 7 |  |
|  | test | averages | 3 |

|  |  |  |  |
| --- | --- | --- | --- |
| A2\_FR\_AVG\_WAHR | A2\_FR\_AVG\_FALSCH | A2\_FR\_AVG\_UNBEKANNT |  |
| 5 | 5 | 2 |  |
|  | test | averages | 4 |

|  |  |  |  |
| --- | --- | --- | --- |
| B1\_EN\_AVG\_WAHR | B1\_EN\_AVG\_FALSCH | B1\_EN\_AVG\_UNBEKANNT |  |
| 6 | 6 | 17 |  |
|  | test | averages | 5 |

|  |  |  |  |
| --- | --- | --- | --- |
| B1\_FR\_AVG\_WAHR | B1\_FR\_AVG\_FALSCH | B1\_FR\_AVG\_UNBEKANNT |  |
| 9 | 9 | 10 |  |
|  | test | averages | 6 |

|  |  |  |  |
| --- | --- | --- | --- |
| B2\_EN\_AVG\_WAHR | B2\_EN\_AVG\_FALSCH | B2\_EN\_AVG\_UNBEKANNT |  |
| 11,75 | 18,5 | 17 |  |
|  | test | averages | 7 |

|  |  |  |  |
| --- | --- | --- | --- |
| B2\_FR\_AVG\_WAHR | B2\_FR\_AVG\_FALSCH | B2\_FR\_AVG\_UNBEKANNT |  |
| 8,75 | 14 | 17,25 |  |
|  | test | averages | 8 |

|  |  |  |  |
| --- | --- | --- | --- |
| C1\_EN\_AVG\_WAHR | C1\_EN\_AVG\_FALSCH | C1\_EN\_AVG\_UNBEKANNT |  |
| 11,6 | 10 | 28,8 |  |
|  | test | averages | 9 |

|  |  |  |  |
| --- | --- | --- | --- |
| C1\_FR\_AVG\_WAHR | C1\_FR\_AVG\_FALSCH | C1\_FR\_AVG\_UNBEKANNT |  |
| 8 | 12,2 | 24,4 |  |
|  | test | averages | 10 |