Köln, 15. Oktober 2018

Universität zu Köln
MA: Informationsverarbeitung
SM1: Verarbeitung textueller Daten
Dozent: Prof. Andreas Witt

Expansion von Morphemkoordinationen zur Verbesserung der Informationsextraktion

vorgelegt von

Johanna Binnewitt
Matrikelnummer 5429692
E-Mail: binnwij@smail.uni-koeln.de
Dellbrücker Hauptstr. 89
51069 Köln

Inhalt

1. Einleitung

Informationsextraktion hilft dabei, relevante Informationen aus unstrukturiertem Text zu ziehen. Treten diese Informationen jedoch in koordinierten Strukturen auf, so ist es bei einem musterbasierten Ansatz keineswegs leicht, die gewünschten Informationen zu identifizieren. Genau dieses Problem entsteht auch bei der Informationsextraktion von Kompetenzen und Arbeitsmitteln aus Stellenausschreibungen. Da es sich bei dieser Art von Text um unstrukturierte Daten handelt, treten häufig Formulierungen auf, in denen Teile von Wörtern scheinbar fehlen. Ein musterbasierter Ansatz kann beispielsweise in dem Ausdruck „Steuer- und Sozialversicherungsrecht“ lediglich das letzte Konjunkt, also „Steuerversicherungsrecht“, als vollständiges Wort identifizieren und „übersieht“ den elidierten Ausdruck „Steuerrecht“. Dadurch gehen viele Informationen verloren, die eigentlich extrahiert werden sollen. Auslöser für das Verschwinden eines Wortteils scheinen identisches Wortmaterial sowie die Konjunktion „und“ zu sein. Dies gilt allerdings nicht nur auf Wortteil-Ebene, sondern kann auch für Wörter und Konstituenten gelten (mehr dazu siehe Kapitel 2 und 3).

Um Informationen aus koordinierten Strukturen auffindbar zu machen, bietet die vorliegende Arbeit einen explorativen Ansatz, koordinierte Strukturen in Stellenausschreibungen so aufzulösen, dass mehr relevanten Informationen extrahiert werden können. Die Implementation beschränkt sich aufgrund des Umfangs zunächst ausschließlich auf die Expansion von Morphemkoordinationen. Da es verschiedene Ansätze gibt, welche koordinative Strukturen aus grammatikalischer Sicht unterschiedlich erklären, werden zunächst zwei Theorien vorgestellt, die Koordinationen einerseits aus dependenzgrammatischer Sicht und anderseits als phonologisches Phänomen beschreiben. Anschließend wird auf den speziellen Fall der Morphemkoordination eingegangen. Im zweiten Teil wird das Framework *quenfo* vorgestellt, in das die Expansion von Morphemkoordinationen eingebettet wurde. Die einzelnen Komponenten, die für diese Expansion notwendig sind, werden erläutert und schließlich diskutiert.

2. Koordination aus dependenzgrammatischer Sicht

Um das Phänomen der Koordination aus dependenzgrammatischer Sicht zu erklären, muss zunächst das Prinzip der Dependenz als zugrundeliegendes System vorgestellt werden. Im Gegensatz zu Konstituentengrammatiken, welche Sätze in Konstituenten zerlegt, um diese dann wiederum zu analysieren, beschäftigen sich Dependenzgrammatiken mit den Abhängigkeiten von Wörtern und Morphemen untereinander. Dependenzgrammatiker gehen davon aus, dass

Elemente andere Elemente im Satz bedingen oder voraussetzen, und untersuchen die Beziehungen dieser Wörter zueinander (Lobin 1993: 9). Ramers beschreibt den Unterschied zwischen Konstituenten- und Dependenzgrammatiken in der Art, zwischen welchen Elementen Relationen gezogen werden. Während Erstere Teil-Ganzes-Relationen aufstellen, fokussieren Letztere Abhängigkeitsbeziehungen zwischen Elementen der Satzstruktur (Ramers 2007: 77).

2.1. Dependenzgrammatiken

Eine Dependenzgrammatik legt fest, welche sprachlichen Elemente anderen Elementen untergeordnet sind. Die phonetische bzw. graphische Verkettung wird dabei zunächst nicht abgebildet, sondern entsteht erst später während der Linearisierung (Lobin 1993: 28). Es geht lediglich darum festzulegen, welche Wörter andere Elemente im Satz fordern. Die Einforderung von Elementen wird auch als Valenz bezeichnet. Da es müßig wäre, für jede mögliche Verbindung von Wörtern Regeln festzustellen, abstrahieren Dependenzgrammatiken, indem sie mithilfe sogenannter Satellitentypen Dependenzen zwischen verschiedenen Wortklassen bündeln (Lobin 1993: 17). So bündelt Lobin in seinem Beispiel die folgenden Konstrukte:

(1) a. Paul fährt *nach* Bonn.

b. Paul fährt *dorthin*.

c. Paul fährt den Berg *hinauf*.

d. Paul fährt *zu* Katrin.

Die unterstrichenen Wortgruppen stellen dabei die Satelliten dar, die in diesem Fall einer Direktivergänzung zum Verb (DIR_V) entsprechen. Ein Satellitentyp kann verschiedene Ausdruckelemente haben, die bestimmen, welche Wortformen an der Spitze des Satelliten stehen dürfen. Im vorliegenden Beispiel sind dies Adverbien und Präpositionen (Lobin 1993: 17f.). Die Valenz eines Elements kann wie folgt notiert werden:

(2) [fahren/V [NOM][DIR]]

Das Verb „fahren“ fordert also Nominativ- und Direktivergänzungen. Das die gesamte Phrase regierende Element – in unserem Beispiel das Verb „fahren“ - wird als Nukleus bezeichnet (Lobin 1993: 20). Regierende Elemente können außerdem fakultative Satellitentypen haben. So kann beim folgenden Satz der unterstrichene Teil weggelassen werden:

(3) Paul fährt schnell zu Katrin.

Fakultative Satellitentypen werden durch ein hochgestelltes ‘+‘ markiert:

(4) [fahren/V [NOM][DIR]$^{+}$[MOD]]

2.2. Koordinationen in Dependenzgrammatiken

Im Falle einer Koordination scheinen die Regeln der Rektion verletzt zu werden, da hier mehr Satelliten als eigentlich gefordert auftreten. Dies soll zunächst an folgendem Beispielsatz gezeigt werden:

(5) Paul isst einen Apfel und Katrin eine Birne.

Das Verb „essen“ sollte hier üblicherweise wie folgt strukturiert sein:

(6) [essen/V [NOM][AKK]]

Jedoch treten mit „Paul“ und „Katrin“ zwei Nominalergänzungen und mit „einen Apfel“ und „eine Birne“ zwei Akkusativergänzungen auf. Auslöser für die Vermehrung von Satelliten scheint der Konjunktor „und“ zu sein. Um dieses Phänomen grammatikalisch zu erklären, gibt es laut Lobin zwei Möglichkeiten. So könnte es einerseits sein, dass die Koordinationssyntax als Teil der allgemeinen Syntax beschrieben werden muss. Andererseits wäre es möglich, Koordinationsstrukturen in einer Art „Metasyntax“ zu verorten (Lobin 2006: 973).

Lobin beschreibt die gedoppelten Wortformen als kontrastierend, wobei das erste Konjunkt einen vollständigen Satz darstellt und im zweiten Konjunkt die jeweils kontrastierenden Elemente auftauchen (Lobin 1993: 111f.). Nimmt man das erste Konjunkt, so könnten die kontrastierenden Elemente des zweiten Konjunkts ihre jeweiligen Partner im ersten Konjunkt ersetzen, ohne, dass die Struktur der Phrase verletzt wird:

(7) Katrin isst eine Birne.

Demnach scheinen „in Kontrast stehende Elemente […] also die gleiche Funktion in der dependenziellen Struktur übernehmen zu können“ (Lobin 2006: 980). Der Satz kann innerhalb der Valenznotation folgendermaßen repräsentiert werden:

(8) [< ∅, und > isst

< [NOM Paul], [NOM Katrin]>

< [AKK einen Apfel], [AKK eine Birne]>]

Die kontrastierenden Elemente werden hier jeweils als Paar in spitzen Klammern dargestellt. Lobin definiert hierbei das erste Konjunkt als erste Projektion und die Ersetzung mit Elementen des zweiten Konjunkts (siehe (7)) als zweite Projektion (Lobin 2006: 980).

Die hier präsentierte Erklärung von Koordinationsstrukturen hilft dabei, die Frage zu beantworten, welcher Kontext für jede Expansion mitgenommen werden muss. Beispielsweise expandiert die Formulierung „Erfahrungen in der Bedienung von Bau- und Landmaschinen“ zu „Erfahrungen in der Bedienung von Baumaschinen“ und „Erfahrungen in der Bedienung von Landmaschinen“, da lediglich „Baumaschinen“ und „Landmaschinen“ als kontrastierende Elemente auftauchen. Da wir zuvor gesehen haben, dass in Kontrast stehende Elemente die gleichen Funktionen innerhalb der dependenziellen Struktur einnehmen können, wissen wir nun, dass „Erfahrungen in der Bedienung von“ als Kontext für das zweite kontrastierende Element gelten kann.

3. Koordination als phonologisches Phänomen

Im Gegensatz zu Lobin erklärt Hartmann Koordinationsstrukturen als phonologisches Phänomen (vgl. Hartmann 2000). Am Beispiel von *Right Node Raising* (RNR) geht sie davon aus, dass die Abhängigkeit zwischen einem Konjunkt und einem davon rechtsstehenden Element nicht von syntaktischer Natur sein kann (Hartmann 2000: 53). Deshalb sucht sie eine Erklärung für dieses Phänomen auf einer anderen Ebene der Grammatik und folgert, dass elidierte – also ausgelassene – Elemente einer Konjunktion ein Phänomen auf der Ebene der phonetischen Form darstellen (Hartmann 2000: 22). Als Auslöser erkennt sie wie Lobin identisches Material in allen Konjunkten. Alte Theorien zum RNR, die Hartmann als *Movement Theory* zusammenfasst, sind der Auffassung, dass identische Elemente von allen Konjunkten bis ans rechte Ende des letzten Konjunkts weitergegeben werden:

(9) $[_{XP\ldots}t_{\alpha}]$ & $[_{XP\ldots}t_{\alpha}]$ α

t_{α} stellt hierbei die Position dar, an der das elidierte Element fehlt, welches hinter dem letzten Konjunkt schließlich angereiht ist. Hartmann ersetzt die Struktur der *Movement Theory* durch folgende:

(10) $[\ldots \phi_{Y}]$ & $[\ldots Y]$

Hier enthält das erste Konjunkt weiterhin die phonetische Matrix von Y, jedoch wird diese auf dem Level der phonetischen Form reduziert. Demnach greifen Koordinationsellipsen die Syntax eines Satzes nicht an, da die grammatikalischen Eigenschaften des elidierten Elements bestehen bleiben (Hartmann 2000: 54). Hartmann bezeichnet den nicht von der Ellipse betroffenen Teil des Konjunkts als Überbleibsel (*remnant*) und das elidierte Element als Ziel (*target*). In einem Satz wie

(11) Peter bringt und Marie holt die Kinder.

stellen „Peter bringt“ und „Maria holt“ die Überbleibsel und „die Kinder“ das Ziel dar (ebd.).

Hartmann begründet ihre Theorie, dass Koordinationsellipsen eine Reduktion der phonetischen Form darstellen, unter anderem damit, dass sich elidierte Elemente nicht immer an die Grenzen von Konstituenten halten. Demnach können auch mehrere oder unvollständige Konstituenten das Ziel einer Koordination darstellen. Dieses Argument gilt auch für das in dieser Arbeit vorgestellte Phänomen, da hier die elidierten Elemente lediglich Morpheme umfassen. Im Folgenden soll deshalb diese Variante der Koordination detaillierter untersucht werden.

4. Morphemkoordinationen

Sobald der elidierte Teil einer Koordination lediglich einen Wortteil umfasst, spricht man von einer Morphemkoordination. In schriftlicher Form wird dieses Morphem durch einen Ergänzungsstrich ersetzt (Clematide 2009: 38). Wie Booji anmerkt tritt dieses Phänomen im Deutschen und Niederländischen häufig im Zusammenhang mit komplexen Wörtern auf, und zwar genau dann, wenn zwei Wörter in derselben Phrase in einem Wortteil identisch sind (Booji 1985: 143). Auf den ersten Blick scheint die Ellipse innerhalb eines Worts das Prinzip der Lexikalischen Integrität zu verletzen, wonach die Syntax keinen Einfluss auf die morphologische Struktur nehmen darf. Deshalb geht Booji wie Hartmann davon aus, dass es sich bei der Morphemkoordination um ein phonologisches Phänomen handelt. Er stellt die Hypothese auf, dass koordinierte Strukturen entstehen, indem widerkehrendes Material entfernt wird. Unter anderem stützt Booji diese Annahme auf die Beobachtung, dass sowohl im Deutschen als auch im Niederländischen Interfixe am kontrastierenden Element stehenbleiben. So entstehen Morphemkoordinationen wie die Folgende:

(12) Beratungs- und Verkaufsgespräche

Hier bleibt das ‘s‘ an „Beratung“ als Überbleibsel des ursprünglichen Kompositums bestehen. Diese Beobachtung erlaubt es uns, die Expansion als Ersetzung des Bindestrichs mit dem elidierten Wortteil zu implementieren, da die morphologischen Eigenschaften des kontrastierenden Elements alle bestehen bleiben.

4.1. Links- und rechtselliptische Formen

Im Rahmen der Morphemkoordination kann sich das elidierte Morphem entweder am linken oder am rechten Rand des Worts befinden. Im STTS-Tagset werden rechtselliptische Formen mit TRUNC markiert. Die Definition dieses Tags lautet wie folgt:

„Mit TRUNC werden Wortteile bezeichnet, die mit einem Bindestrich enden, der einen Teil des nachfolgenden, mit *und, oder* verknüpften Wortes ersetzt.“ (Clematide 2009, 40)

Dieses Tag macht es einfach, rechtselliptische Formen im Korpus zu erkennen, da alle von Ellipsen betroffenen Wörter als solche markiert sind. Beispielsweise liefert ein POS-Tagger folgendes Ergebnis:

(13)	Kooperations-	und	Kommunikationsfähigkeit
	TRUNC	KON	NN

Clematide erklärt die Verwendung eines gesonderten Tags damit, dass aus dem übrig gebliebenen Wortteil keine morphologischen Merkmale abgelesen werden können, da diese im Deutschen oft im Suffix enthalten sind (Clematide 2009, 41). Hier kann lediglich das letzte Konjunkt, welches das elidierte Morphem enthält, Aufschluss über die Wortform geben.

Im Gegensatz dazu können bei Linksellipsen aus allen Konjunkten die Wortformen abgelesen werden, da hier das Suffix für jedes Element bestehen bleibt:

(14)	Kooperationsfähigkeit	und	-bereitschaft
	NN	KON	NN

Dies macht es gleichzeitig schwieriger linkselliptische Elemente anhand der POS-Tags im Korpus zu identifizieren. Vielmehr kann hier der anführende Bindestrich als Indiz wirken. Da Linksellipsen jedoch in den Stellenausschreibungen vergleichsweise selten auftreten[1], behandelt die nachfolgende Arbeit ausschließlich Rechtsellipsen.

4.2. Morphemzerlegung

Eine Herausforderung, die bei der Expansion von Morphemkoordinationen entsteht, ist zu erkennen, welcher Wortteil elidiert wurde. Bei einer Rechtsellipse befindet sich der Wortteil, der bei allen anderen Konjunkten fehlt, am rechten Rand des einzigen vollständig realisierten Konjunkts. Jedoch kann ein Parser hier nicht auf Anhieb erkennen, wo er das Konjunkt zerlegen soll. Deshalb gehen wir zunächst auf die Eigenschaften dieses elidierten Teils ein. Booji wirft ebenfalls die Frage auf, welche Beschaffenheit die entfernte Komponente haben soll. Er stellt fest, dass Ellipsen sowohl am linken als auch am rechten Rand eines Worts, jedoch immer

[1] Während in lediglich 31 Abschnitten der Stellenausschreibungen „und -“ auftritt, enthalten 312 Abschnitte „-und“.

unmittelbar neben der Konjunktion stehen. Er prüft zunächst, ob es sich um eine phonologische Komponente handelt. Dies kann jedoch nicht zutreffen, da in diesem Fall auch Koordinationen wie „Vögel und Spiegel“ möglich sein müssten (Booji 1985: 148). Ferner kann der elidierte Teil ebenso wenig morphologischer Natur sein, da in diesem Fall auch Suffixe wie in „Salzig und Mehlig“ entfernt werden würden. Daher führt Booji den Begriff „phonologische Wörter“ ein, mit dem er zum Ausdruck bringen möchte, dass syntaktische Wörter nicht immer mit ihren phonologischen Korrelaten korrespondieren (Booji 1985: 149). Demnach können sowohl Komposita als auch Derivate koordiniert werden. Komposita sind hier so definiert, dass ein Morphem oder Morphem-Komplex aus mindestens zwei lexematischen Morphemen bestehen muss (Höhle 2018: 225). So bilden beide Bestandteile des Kompositums wieder eigenständige Wörter. Ein Beispiel für eine solche Koordination wäre „Deutsch- und Englischkenntnis“, denn sowohl „Englisch“ als auch „Kenntnis“ können als eigenständige Wörter angesehen werden. Jedoch betreffen Morphemkoordinationen auch Derivate wie beispielsweise „be- und entladen“. Hier stellt nur das Morphem „laden“ ein lexematisches Morphem dar, weshalb es laut Höhle als Derivat definiert wird (ebd.).

Des Weiteren muss definiert werden wie mit Konjunkten umgegangen wird, die aus mehr als zwei Wortteilen bestehen, da hier unklar ist welche der Wortteile zum kontrastierenden und welche zum elidierten Element gehören. Als Beispiel dafür dient folgende koordinierte Struktur:

(15) Alten- und Krankenfachpflegekraft

Beim zweiten Konjunkt besteht der kontrastierende Teil lediglich aus dem Morphem „Kranken“. Demnach kann für die Identifikation des elidierten Elements keine Regel verwendet werden, die stets den letzten Wortteil des vollständigen Konjunkts als elidierten Teil identifiziert.

5. Anwendungsfall: IE aus Stellenausschreibungen

Die Beachtung koordinativer Strukturen kann unter anderem bei der Informationsextraktion von Bedeutung sein. Im Folgenden wird daher gezeigt, wie die Auflösung von Morphemkoordinationen dazu beiträgt die Extraktion von Kompetenzen und Arbeitsmitteln aus Stellenausschreibungen zu verbessern. Für die allgemeine Informationsextraktion liegt mit *quenfo*[2] bereits ein Framework vor, welches unter anderem dafür entwickelt wurde, aus einem Korpus von Stellenausschreibungen die geforderten Kompetenzen sowie die verwendeten Arbeitsmittel zu extrahieren. Dabei wird ein Bootstrapping-Ansatz verwendet, der mithilfe von bekannten Mustern relevante Informationseinheiten sowie neue Muster entdeckt (vgl. Geduldig 2017). Diese relevanten Informationseinheiten sollen der Qualifikationsentwicklungsforschung dienen, sodass Entwicklungen auf dem Arbeitsmarkt hinsichtlich geforderter Kompetenzen sowie verwendeter Arbeitsmittel besser reguliert werden können. Jedoch enthalten die extrahierten Einheiten oft koordinierte Strukturen, welche es beispielsweise unmöglich machen, die Kompetenz „Deutschkenntnisse“ aus „Deutsch- und Englischkenntnisse“ zu extrahieren. Aus diesem Grund wurde im Rahmen der vorliegenden Arbeit das Framework erweitert, sodass koordinierte Informationseinheiten aufgelöst und so mehr Informationen gewonnen werden können.

Da es sich bei den Daten um Stellenausschreibungen handelt, ist zunächst festzuhalten, dass es sich um ein domänenspezifisches Vokabular handelt. Dies bringt Vorteile für die Implementierung der Morphemzerlegung mit sich, denn die expandierten Wörter sind aller Wahrscheinlichkeit nach bereits im Korpus aufgetreten und die richtige Zusammensetzung kann demnach mit einer Wörterliste abgeglichen werden (mehr dazu unter 5.4). Durch diesen regelbasierten Ansatz können bereits viele Fälle von Expansionen abgedeckt werden.

5.1. Testumgebung zur Überprüfung der Koordinationsauflösung

Um die Auflösung der Koordinationen gesondert betrachten zu können, wurde ein *JUnit*-Test geschrieben[3], welcher bereits extrahierte Kompetenzen bzw. Arbeitsmitteln einliest und darin enthaltene Koordinationen auflöst. Für beide Informationsarten liegt jeweils eine Datenbank vor, welche neben anderen Spalten folgende Angaben enthält:

[2] https://github.com/spinfo/quenfo
[3] Installationsanleitung: siehe Anhang 7.1.

Tabelle 1: Auszug aus der Kompetenzdatenbank

Sentence	Comp	CompResolved
fundierte Fachkenntnisse in den speziellen Aufgabenstellungen der Energiewirtschaft rund um die Zähl- und Messtechnik	zähl und messtechnik	zähltechnik; messtechnik
Als idealer Bewerber verfügen Sie über ein abgeschlossenes Studium der Elektrotechnik oder eine vergleichbare Ausbildung zum Techniker und bringen gute Kenntnisse in der Energie- und Antriebstechnik mit.	energie und antriebstechnik	energietechnik; antriebstechnik

Die Spalte „Comp“ (bzw. „Tool“) enthält dabei die bereits extrahierte Informationseinheit, die Spalte „CompResolved“ (bzw. „ToolResolved“) gibt an wie die koordinierten Ausdrücke expandiert werden sollen. Die Abkoppelung von der zuvor erfolgten Informationsextraktion ermöglicht es, die Koordinationsauflösung separat zu betrachten. Somit liefert diese Arbeit auch keine Evaluation der Informationsextraktion. In *quenfo* werden die extrahierten Einheiten als Lemmata gespeichert. Dadurch sind die identifizierten Informationseinheiten unabhängig von der Wortform, sodass eine anschließende Kategorisierung besser durchgeführt werden kann. Aufgabe ist es nun, koordinierte Strukturen in den Informationseinheiten zu entdecken, diese aufzulösen und die expandierten Informationseinheiten zu exportieren. Im Folgenden soll erläutert werden, wie die Erkennung und die Auflösung von Morphemkoordinationen sowie die Morphemzerlegung des vollständigen Konjunkts implementiert wurden. Der Export aller aufgelösten Informationseinheiten erfolgt schließlich in eine Datenbank.

5.2. Erkennung von Morphemkoordinationen

Wie bereits erwähnt enthalten Informationseinheiten oft koordinierte Strukturen, da die Muster zur Informationsextraktion so gewählt wurden, dass auch koordinierte Strukturen extrahiert werden. Dadurch bietet sich die Möglichkeit, nur diese Koordinationen aufzulösen. Zwar wäre es möglich, die gesamte Stellenausschreibung auf koordinierte Strukturen zu untersuchen und diese aufzulösen – dies hätte den Vorteil, dass dadurch noch unentdeckte relevante Informationen zutage kommen könnten -, jedoch wäre dies mit einer höheren Laufzeit verbunden. Deshalb wird eine Koordination nur ausgelöst, wenn sich in der Informationseinheit

eine Konjunktion befindet. Informationseinheiten mit einer Konjunktion werden an den `CoordinationResolver` übergeben. Dieser prüft zunächst, ob sich unter den Wortarten der Koordination ein „TRUNC“ befindet. Dies zeigt an, dass es sich um eine rechtselliptische Koordination handelt. Da sich im Korpus in den Informationseinheiten keine linkselliptischen Strukturen befinden, wird im Folgenden nur auf die Auflösung von Rechtsellipsen eingegangen.

5.3. Expansion von Morphemkoordinationen

Die erste Herausforderung besteht darin, herauszufinden, welche Wortart koordiniert wurde. Diese Information hilft dabei, das Ende der koordinierten Struktur zu bestimmen. Beispielsweise werden im nachfolgenden Ausdruck die Adjektive koordiniert, wobei das nachfolgende Nomen den Kontext innerhalb der Informationseinheit darstellt:

(16) kommunikations- und kooperationsbereite Einstellung

Somit besteht der zu expandierende Teil lediglich aus „kommunikations- und kooperationsbereite“. Da die elliptischen Elemente nur als „TRUNC“ markiert sind, lässt sich über das Tag nicht die Wortart des letzten Konjunkts feststellen. Daher wird die Wortart der kontrastierenden Elemente darüber bestimmt, ob das erste Token der Koordination klein- oder großgeschrieben ist. Handelt es sich um ein großgeschriebenes Token, so wird angenommen, dass es ein Nomen ist. In diesem Fall wird überprüft, ob vor dem Nomen möglicherweise noch ein Modifikator steht. Anschließend wird die Endposition der Koordination mithilfe der ermittelten koordinierten Wortart bestimmt. Hier kann der Fall eintreten, dass die zuvor identifizierte Informationseinheit nicht die gesamte koordinierte Struktur umfasst. Beispielsweise besteht eine Informationseinheit in der Kompetenzdatenbank aus „Deutsch- und gute“, wobei „Englischkenntnisse“ abgeschnitten ist. In diesem Fall wir für „NN“ in der Informationseinheit kein Index gefunden, sodass die Koordination vervollständigt werden muss bis ein „NN“ enthalten ist. Da nun der gesamte Abschnitt der Koordination bekannt ist, müssen alle Konjunkte mitsamt ihrer Modifikatoren gesammelt werden. Hier muss überprüft werden, ob eine koordinierte Einheit unmittelbar vor sich einen Modifikator mitführt. Ist dies der Fall, so wird der Modifikator zum koordinierten Element hinzugefügt. Sollte es keinen Modifikator geben, so wird der Modifikator des vorherigen Konjunkts übernommen:

(17)
```
gute Deutsch-, sehr gute Französisch- und Englischkenntnisse
gute - Deutsch-
sehr gute - Französisch-
sehr gute - Englischkenntnisse
```

Die Konjunkte werden mitsamt ihrer Modifikatoren anschließend an die Methode `combineCoordinations` übergeben. Diese Methode übernimmt die tatsächliche Koordinationsauflösung. Das letzte Konjunkt wird genutzt, um den elidierten Teil der Koordinationsellipsen zu ermitteln (siehe 5.4), welcher dann an alle vorherigen Konjunkte angehängt wird. Dabei wird der Bindestrich durch das ermittelte Morphem ersetzt, sodass beispielsweise „Deutsch-“ zu „Deutschkenntnisse“ wird. An dieser Stelle wird deutlich, warum es sinnvoll ist, die Expansion auf Ebene der Tokens und nicht der Lemmata durchzuführen. Bei einem Lemma wären mögliche Interfixe am kontrastierenden Element bereits entfernt worden, die an dieser Stelle keine einfache Ersetzung des Bindestrichs mit dem elidierten Material möglich machen würden. Einen Sonderfall für die Zusammensetzung der Konjunkte stellen Wörter dar, die ursprünglich auch mit Bindestrich geschrieben werden (z.B. „SAP-Kenntnisse“). Hier gilt die Regel, dass der Bindestrich bestehen bleibt und das Morphem nach dem Bindestrich großgeschrieben wird, sobald das „TRUNC“-Token nur aus Großbuchstaben besteht.

Die Rückgabe der Methode `combineCoordinations` liefert beispielsweise für die Eingabe („gute Deutsch-“, „sehr gute Französisch-“, „sehr gute Englischkenntnisse“) die Ausgabe („gute Deutschkenntnisse“, „sehr gute Französischkenntnisse“, „sehr gute Englischkenntnisse“). Diese Alternativen werden nun in den übriggebliebenen Kontext der Informationseinheit eingebettet und anschließend erneut lemmatisiert. Dies ist notwendig, da die vom *quenfo* extrahierten Informationseinheiten ebenfalls lemmatisiert sind und somit die Vergleichbarkeit gewährleistet sein muss. Besteht die Informationseinheit jedoch lediglich aus einem Token, so schlägt die Lemmatisierung in manchen Fällen fehl. Um fehlerhafte Lemmatisierungen aufzudecken, wird das Lemma des letzten Konjunkts zu Rate gezogen. Stimmen das Suffix dieses Lemmas und das Suffix des neuen Lemmas nicht überein, so wird das neue Lemma an das Lemma des letzten Konjunkts angeglichen.

5.4. Morphemzerlegung

Neben der Ermittlung der koordinierten Tokens ist die Morphemzerlegung des letzten Konjunkts ein wichtiger Schritt zur Koordinationsexpansion, da hier der Substring ermittelt wird, mit dem alle unvollständigen Konjunkte kombiniert werden. Dafür wird in der Methode `combineCoordinations` die externe Bibliothek `JWordSplitter`[4] verwendet, welche ein Kompositum in sämtliche Bestandteile zerlegen kann. Üblicherweise zerlegt der

[4] https://github.com/danielnaber/jwordsplitter (zuletzt aufgerufen am 15.10.18)

`AbstractWordSplitter` Komposita in lexikalische Morpheme, jedoch können ihm auch manuell Ausnahmen übergeben werden. Dies ist vor allem hilfreich, da Präfixe wie „be“ in „beladen“ regulär nicht abgetrennt werden. Ausnahmen, die für die vorliegende Domäne der Stellenausschreibungen gelten, werden in einer .txt-Datei[5] gespeichert, wobei die gewünschte Trennung des Worts durch einen senkrechten Strich markiert wird (z.B. „Alten|fachpflegekraft“).

Falls keine Ausnahme für das zu zerlegende Konjunkt vorliegt, greifen die vom Splitter definierten Regeln für die Kompositazerlegung. Entstehen dabei zwei Morpheme, so wird das letzte Morphem als Suffix für alle weiteren Konjunkte gewählt. Sobald bei der Zerlegung mehr als zwei Teile entstehen, wird das letzte Morphem als Suffix definiert. Da diese Herangehensweise heuristisch ist, wird die entstandene Trennung des letzten Konjunkts in eine Evaluationsdatei[6] übernommen, damit der Nutzer die Trennungsstelle gegebenenfalls korrigieren kann. Die korrigierten Zerlegungen werden beim nächsten Durchlauf des Programms in die Ausnahmen-Datei übernommen. Das ermittelte elidierte Element wird wie in 5.3. beschrieben mit allen kontrastierenden Teilen zusammengefügt, indem es den Bindestrich ersetzt. Schließlich werden alle entdeckten Konjunkte sowie die Koordination als Ganzes in eine Datenbank[7] exportiert.

6. Diskussion

Die vorliegende Arbeit sollte explorativ zeigen, welche Phänomene im Rahmen der Morphemkoordination auftreten können und wie diese bei der Expansion berücksichtigt werden müssen. Nachdem theoretische Ansichten über die Struktur von Koordinationen präsentiert wurden, ging es im praktischen Teil darum, die Erkenntnisse aus der Theorie im Framework umzusetzen. Eine quantitative Evaluation ist im Rahmen dieser Arbeit nicht möglich, da zu wenig Trainingsdaten vorliegen, um Aussagen über die Güte des Systems zu treffen. Viel mehr wurde das Framework sukzessiv an die aus den Informationseinheiten bekannten Phänomene angepasst. Trotzdem kann das vorgestellte Framework weiterhin an vielen Stellen verbessert werden.

Da im Korpus Rechtsellipsen dominieren, wurde die Expansion zunächst nur auf diese Art der Koordination angepasst. Jedoch sollte das Framework in Zukunft um die Behandlung von

[5] Pfad: src/test/resources/coordinations/resolvedCompounds.txt
[6] Pfad: src/test/resources/coordinations/possibleCompounds.txt
[7] Pfad: src/test/resources/coordinations/Coordinations_Competences_2011.db bzw. Coordinations_Tools_2011.db

Linksellipsen erweitert werden. Des Weiteren könnten Informationseinheiten ohne Konjunktion darauf überprüft werden, ob sie nicht doch zu einer koordinierten Struktur gehören und dementsprechend expandiert werden können. Das gleiche gilt für abgeschnittene Modifikatoren, die sich außerhalb der Informationseinheit befinden. Außerdem könnte in Bezug auf die Morphemzerlegung ein Vollformenlexikon aus dem gesamten Korpus der Stellenausschreibungen erstellt werden. Dies könnte sich den Vorteil zu Nutzen machen, dass von Ellipsen betroffene Wörter mit hoher Wahrscheinlichkeit vollständig realisiert im Korpus auftauchen. Dadurch könnte der elidierte Teil einer Morphemkoordination leichter identifiziert werden. Zuletzt könnte die Semantik des Konjunktors ebenfalls berücksichtigt werden. Bisher ist die Anwendung so implementiert, dass „Deutsch- und Englischkenntnisse“ sowie „Deutsch- oder Englischkenntnisse“ dasselbe Ergebnis liefern. Eine Berücksichtigung des logischen Operators würde allerdings eine umfassende Änderung der Abläufe in *quenfo* fordern, da nicht-koordinierte Informationseinheiten ebenfalls von logischen Operatoren betroffen sein können.

7. Anhang

7.1. Installationsanleitung

quenfo ist in Java geschrieben und ein Maven-Projekt. Die Implementation wurde in Java 1.7 und auf einem Windows10-Betriebssystem entwickelt. Das Github-Repository[8] kann innerhalb von Git oder als Zip-Datei heruntergeladen werden. Als Entwicklungsumgebung wurde Eclipse Photon (4.8.0) verwendet. Die Stellenausschreibungen und die expandierten Informationseinheiten werden jeweils in .db-Dateien gespeichert. Zum Betrachten dieser Dateien kann beispielsweise der SQLite-Browser[9] genutzt werden.

Der *JUnit*-Test `EvaluateCooResolvation`[10] führt den in dieser Arbeit beschriebenen Ablauf aus. Die beigefügte Zip-Datei enthält alle notwenigen Daten, die unter „src/test/resources/coordinations“ eingefügt werden sollen:

- „CorrectableCompetences_2011_Gold“: alle Informationseinheiten, die Kompetenzen enthalten, sowie expandierte Informationseinheiten, falls Konjunktion enthalten ist
- „CorrectableTools_2011_Gold“: alle Informationseinheiten, die Arbeitsmittel enthalten, sowie expandierte Informationseinheiten, falls Konjunktion enthalten ist
- „possibleCompounds“: enthält Morphemzerlegungen, die vom System selbst entwickelt wurden
- „resolvedCompounds“: enthält Morphemzerlegungen, die der Splitter als Ausnahmen behandeln soll

Je nachdem, ob Kompetenzen oder Arbeitsmittel expandiert werden sollen, müssen innerhalb des *JUnit*-Tests die Attribute `type`, `inputDB` und `outputDB` angepasst werden.

Ferner benötigen die verwendeten NLP-Tools Modelle, die wie folgt eingefügt werden sollen:

- de-sent.bin & de-token.bin[11]: „information_extraction/data/openNLPmodels/”
- ger-tagger+lemmatizer+morphology+graph-based-3.6+.tgz[12] (als Ordner entpackt): „information_extraction/data/sentencedata_models/”

[8] https://github.com/spinfo/quenfo (zuletzt aufgerufen am 15.10.18)
[9] https://sqlitebrowser.org/ (zuletzt aufgerufen am 15.10.18)
[10] Pfad: src/test/java/quenfo/EvaluateCooResolvation.java
[11] http://opennlp.sourceforge.net/models-1.5/ (zuletzt aufgerufen am 15.10.18)
[12] https://code.google.com/archive/p/mate-tools/downloads (zuletzt aufgerufen am 15.10.18)

8. Literaturverzeichnis

Booij, Geert. (1985) “Coordination Reduction in Complex Words: A Case for Prosodic Phonology”. in Hulst and Smith (eds.) (1985) Advances in Nonlinear Phonology 143 - 160.

Clematide, Simon R. (2009) *Koordination im Deutschen und ihre syntaktische Desambiguierung*. University of Zurich, Faculty of Arts.

Geduldig, Alena (2017) *Muster und Musterbildungsverfahren für domänenspezifische Informationsextraktion*. Masterarbeit. URL: http://www.spinfo.phil-fak.uni-koeln.de/sites/spinfo/arbeiten/Masterthesis_Alena.pdf

Hartmann, Katharina (2000) *Right Node Raising and Gapping: Interface Conditions on Prosodic Deletion*. Philadelphia, PA: John Benjamins Publishing Co.

Höhle, Tilman (2018) “On composition and derivation: The constituent structure of secondary words in German” in Stefan Müller, Marga Reis & Frank Richter (eds.) (2018) *Beiträge zur deutschen Grammatik: Gesammelte Schriften von Tilman N. Höhle*. Berlin: Language Science Press.

Lobin, Henning (1993) *Koordinationssyntax als prozedurales Phänomen*. Tübingen: Gunter Narr Verlag.

– (2006) „Koordination in Dependenzgrammatiken“. in Ágel, Vilmos et all. (eds.) (2006) *Dependenz und Valenz: Ein internationales Handbuch der zeitgenössischen Forschung*, Berlin, New York: Walter de Gruyter https://doi.org/10.1515/9783110171525.2.7.973 .

Ramers, Karl Heinz (2007) *Einführung in die Syntax*. Paderborn: Walter Fink Verlag.