Universität zu Köln
Institut für Linguistik
Abteilung Sprachliche Informationsverarbeitung

Hausarbeit

Suffix Trees und Sprachen mit freier Wortstellung

Hauptseminar: Angewandte Linguistische Datenverarbeitung

WiSe 2009/10

Dozent: Prof. Dr. Jürgen Rolshoven

Oskar Henry Solich

Inhaltsverzeichnis

1 Einleitung

Bei Suffix Trees handelt es sich um Datenstrukturen, die eine Zeichenfolge mit allen Suffixen in komprimierter Form speichern. Mit Suffix werden Teilfolgen einer Zeichenfolge bezeichnet, die am Ende dieser Zeichenfolge steht. Bei der Zeichenfolge „Vier“ sind dies beispielsweise „ier“, „er“ und „r“. Diese Suffix Trees werden vor allem in der Bioinformatik zur Verarbeitung von Gen-Sequenzen sowie in der elektronischen Sprachverarbeitung beim Information Retrievel und für automatische Übersetzungen verwendet. Gerade bei der Sprachverarbeitung zeigen sich jedoch Probleme bei Sprachen, die keine starre Wortstellungsstruktur aufweisen. Der Aufbau der Suffix Trees, deren Anwendungsgebiete in der Sprachverarbeitung, sowie die auftretenden Probleme bezüglich Sprachen mit freier Wortstellung im Satz und mögliche Lösungsansätze hierzu, werden in der vorliegenden Arbeit erörtert, die als Grundlage für die Entwicklung funktionstüchtiger Lösungen dienen soll.

2 Suffix Trees in der Sprachverarbeitung

2.1 Verwendungsgebiete von Suffix Trees

Suffix Trees ermöglichen die Suche von Zeichenketten, sogenannten Strings innerhalb eines Textes in linearer Zeit. Dadurch können Suchstrings besonders schnell gefunden und deren Häufigkeit ermittelt werden. Auch kann die Kontextabhängigkeit von Wörtern und Morphemen ermittelt werden, wodurch sich eine Struktur des Textes konstruieren lässt.

Vorteile bringen Suffix Trees beispielsweise beim Vergleich von Texten, da Ähnlichkeiten in Sätzen durch das einfache Auffinden gemeinsamer Sequenzen leichter und schneller gefunden werden können, was für die Rekonstruktion von Wörtern, Wortarten und Flexionen und damit für die Satzbedeutung entscheidend sein kann. Das Gleiche gilt auch für automatische Textzusammenfassungen, bei denen zuvor die wichtigen Bestandteile ermittelt werden müssen, etwa indem häufig vorkommende Sequenzen zusammengezogen werden. Eine der häufigsten Anwendungen für den Suffix Tree ist das Information Retrieval, also das „computergestützte[...] Suchen nach komplexen Inhalten“ (Wikipedia IR). Hierbei wird nicht nach bestimmten Stichwörtern gesucht, sondern nach Bereichen, die eher schwer abgrenzbar sind und sich somit schlecht in präzisen Suchbegriffen ausdrücken lassen. Durch Suffix Trees lassen sich Verbindungen zwischen verschiedenen Strings ermitteln, wodurch es möglich ist, „Bedeutungswolken“ zu bilden. Das heißt, die Bedeutung eines Wortes kann mit einem anderen Wort in Zusammenhang gebracht und damit dem gleichen Bedeutungsspektrum zugeordnet werden.

Das wiederum ist auch für elektronische Übersetzungen anwendbar, in denen Phrasen, Wörter oder Morpheme[1] der unterschiedlichen Sprachen einander zugeordnet werden müssen, um die Bedeutung des Quelltextes in der Ausgangssprache in den Zieltext der Zielsprache übertragen zu können.

---

1 *Phrasen* sind Einheiten zusammengehörender Wörter innerhalb eines Satzes, bestehend aus mindestens einem Wort. Eine Phrase kann in der Regel nicht durch eine andere Phrase unterbrochen werden. Dies gilt allerdings nur bedingt, wie weiter unten beschrieben.
Morpheme sind die kleinsten bedeutungstragenden Einheiten, also Wortbestandteile, die mindestens eine lexikalische oder grammatikalische Bedeutung tragen. Morpheme, die scheinbar keine Bedeutung tragen, wie etwa die Fugenmorpheme (z. B. Schwein*s*hachse) in deutschen Komposita (Wortzusammensetzungen), haben oft eine phonologische bzw. sprachökonomische Bedeutung.

2.2 Aufbau von Suffix Trees

„A suffix tree is a trie-like data structure representing all suffixes of a string.“

(Ukkonen 1995, 2)

Der Begriff 'Trie' „leitete sich aus dem englischen Ausdruck Information Retrieval ab“ (Wikipedia T). Allgemein betrachtet ist ein Trie lediglich eine Datenstruktur, die Strings in einer Baumstruktur speichert. Die Kanten des Baums repräsentieren dabei die einzelnen Zeichen. In erster Linie wird ein Trie zur Suche in Fließtexten, also vor allem beim Information Retrieval, genutzt, wobei die Texte zuvor in eben diesem Trie gespeichert wurden. Als Suffix Trie wird ein Trie bezeichnet, der eine einzelne Zeichenkette mit allen Suffixen speichert (siehe Abb. 1). Das Ende eines Strings wird dabei als eigenes Zeichen interpretiert, um bei der Verarbeitung von mehreren Strings innerhalb eines Trie diese hinterher wieder auslesen zu können. Das Stringendezeichen kann durch jedes Zeichen realisiert werden, das nicht zum zu verarbeitenden Alphabet gehört. In Abbildung (1) ist dies beispielsweise das Dollarzeichen.

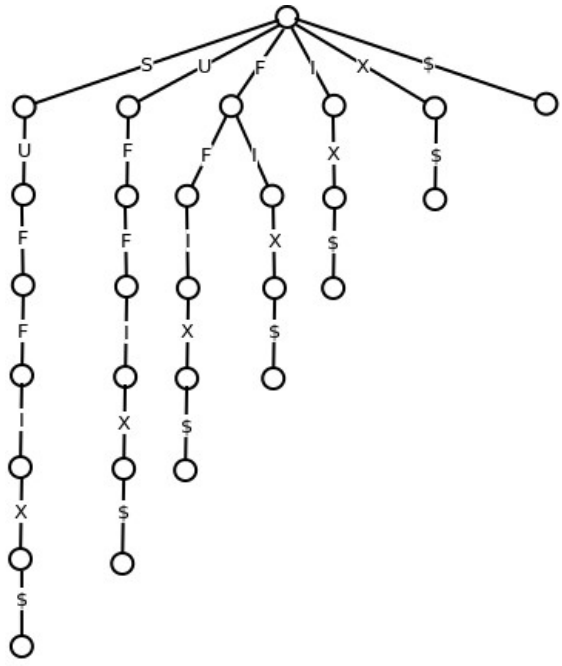

Abb. 1: Suffix Trie

Die komprimierte Variante des Tries ist der Patricia-Trie. Der Unterschied zum unkomprimierten Trie besteht darin, dass der Patricia-Trie nur Knoten enthält, die mindestens

zwei Kindknoten aufweisen oder Blätter[2] sind. Dies wird dadurch erreicht, dass Zeichen, die in einem unkomprimierten Trie jeweils eine eigene Kante zu einem einzelnen Kindknoten bilden mit der darunter liegenden Kante (rekursiv) zusammengefasst werden. Abbildung (1) kann dabei als Beispiel für einen unkomprimierten Trie dienen, während Abbildung (2) einen Patricia-Trie zeigt.

Wie in der Einleitung bereits erwähnt, sind Suffix Trees Datenstrukturen, die eine Zeichenfolge mit allen Suffixen speichern. Im Gegensatz zum Suffix Trie handelt es sich bei dem Suffix Tree um einen Patricia-Trie, also einen komprimierten Trie.

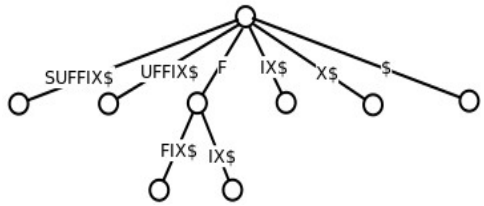

Abb. 2: Suffix Tree

Der Vorteil einer solchen Komprimierung besteht darin, dass Zeit und Speicherplatz bei der Erstellung eingespart werden können. Während der Suffix Trie O(n^2) an Zeit und Speicher benötigt, ist es bei einem Suffix Tree gerade einmal O(n), wenn n die Länge des Strings darstellt.[3] Die Suffix Trees können also mit einem linearen Aufwand erstellt werden, während der Aufwand bei Suffix Tries quadratisch steigt. Die Anzahl der Knoten in einem Suffix Tree kann dabei nicht über $2n$ steigen. (vgl. Nelson 1996)

Einen sehr einfachen und anschaulichen Algorithmus hat Ukkonen in seinem Artikel „On-line construction of suffix trees“ 1995 vorgestellt. Es handelt sich dabei laut Nelson (1996) um die Weiterentwicklung des Algorithmus, den McCreight zwanzig Jahre zuvor entwickelt hat. Der Ukkonen-Algorithmus beginnt mit einem Knoten ohne Nachbarn. Anschließend wird der Text von links nach rechts zeichenweise in den Baum eingesetzt. Beim Beispiel des Strings „SUFFIX“ wird zunächst das „S“ an die Kante eines neuen Knotens gesetzt. Dann wird für „U“ ein weiterer Knoten erzeugt und das „U“ zusätzlich an das „S“ auf der ersten Kante angehängt. Das Gleiche geschieht mit dem ersten „F“. Zuerst wird ein Knoten erstellt, dann

2 Mit *Blättern* werden in Bäumen Knoten bezeichnet, die nur ein Elternknoten haben. Sie sind also kindlos.

3 *O* ist ein Bachmann-Landau-Symbol. Diese Symbole „werden […] bei der Analyse von Algorithmen verwendet und geben ein Maß für die Anzahl der Elementarschritte in Abhängigkeit von der Größe der Eingangsvariablen an.“ (Wikipedia LS)

das „F“ an das „U“ und an das „SU“ angehängt. (vgl. dazu Abb. 3)

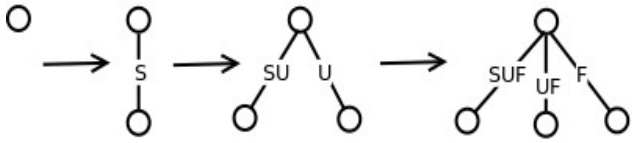

Abb. 3: Konstruktion eines Suffix Trees

Beim zweiten „F“ wird kein neuer Knoten erstellt, da das Zeichen bereits vorhanden ist. Statt dessen wird dieser Schritt übersprungen und das „F“ lediglich an die bereits vorhandenen Knoten angehängt. Beim nächsten Zeichen, dem „I“ gibt es nun zwei Möglichkeiten, nämlich eine Zeichenfolge „FI“ und eine Zeichenfolge „FFI“. Hier muss also noch ein zusätzlicher Knoten eingefügt und die Zeichenkette „FF“ in zwei Zeichenketten getrennt werden, sodass am Wurzelknoten die Zeichenkette „F“ hängt und am darunter liegenden Knoten die Zeichenketten „FI“ und „I“. (Vgl. dazu Abb. 4)

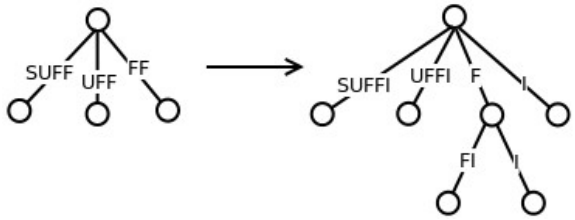

Abb. 4: Spaltung des Knotens bei "F"

Der String wird so bis zum Ende durchlaufen, wodurch anschließend alle Suffixe des Strings im Baum enthalten sind. Der vollständige Suffix Tree zu diesem Beispiel ist Abbildung (1).

Martin Kay (2003 u. 2004) hat beschrieben, wie der Suffix Tree nochmals komprimiert werden kann, indem nur noch Positionsangaben im Baum gespeichert werden. Dadurch wird die Länge der Kanten auf ein Zahlenpaar beschränkt. Die Kanten haben also immer eine konstante Länge und der Baum braucht insgesamt weniger Speicherplatz. Ein String der Länge n (n-1 reguläre Zeichen und das Stringendezeichen) hat n+1 Positionsangaben von 0 bis n. Zum Beispiel hat der String „Anastasia“ eine Länge von 10 (9 reguläre Zeichen und das Stringendezeichen) und somit 11 Positionsangaben von 0 bis 10:

(1) $_0A_1N_2A_3S_4T_5A_6S_7I_8A_9\$_{10}$

Das Zahlenpaar <0,10> beispielsweise bezeichnet hier den kompletten String, während etwa <2,5> für den Teilstring „AST“ steht. Ein Beispiel, wie das im Baum aussieht ist in Abbildung (5) gegeben.

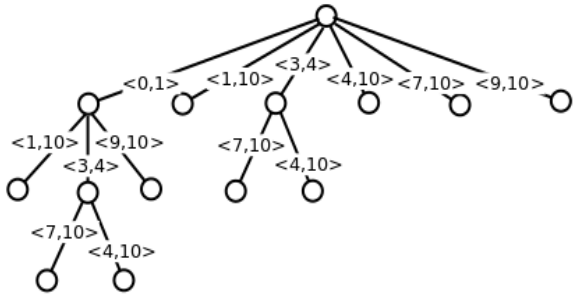

Abb. 5: Suffix Tree von "ANASTASIA" nach Kay (2004)

Die Zahlenpaare sind immer die kleinste Möglichkeit, geben also zunächst nur die erste Position eines Teilstrings an. Dass dies ausreichend ist, zeige ich weiter unten.

Sozusagen eine dritte Stufe der Komprimierung ist der acyclische Wortgraph, bei dem gleiche Teilbäume vereinigt werden. Das heißt, ein Teilbaum, der mehrmals vorkommt, braucht nur einmal zu existieren. Die Kanten, die zu einzelnen gleichen Teilbäumen führen würden, werden so umgestellt, dass sie zu dem selben Teilbaum führen. Der Wurzelknoten des Teilbaums hat somit mehrere Elternknoten, wodurch der gesamte Graph zwar noch gerichtet, aber kein Baum mehr ist:

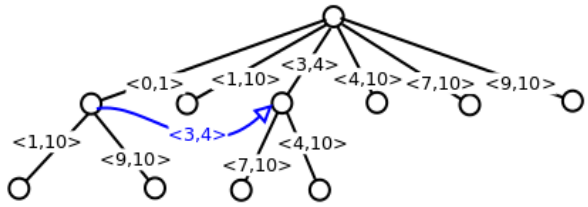

Abb. 6: acyclischer Wortgraph

Die Umsetzung der Suche nach Suchmustern in den komprimierten Suffix Trees und den acyclischen Wortgraphen ist sehr einfach, da nun keine Stringverarbeitung im eigentlichen Sinne stattfindet, sondern die Ergebnisse sozusagen berechnet werden können. Ob dabei nun ein Suffix Tree oder ein acyclischer Wortgraph vorliegt macht keinen Unterschied.

Soll etwa das Suchmuster „AS“ im Text „ANASTASIA“ gesucht werden, wird zunächst überprüft, wo das erste „A“ ist. Es befindet sich an Stelle <0,1>. Nun kann im Suffix Tree nachgesehen werden, an welcher Stelle der Text mit „S“ weiter geht (Position <3,4>). Die

Anfangsposition eines Suchstrings der Länge *n*, dessen Endposition *e* bekannt ist, lässt sich durch Abzug der Länge von der Endposition (*e*-*n*) bestimmen. Im Beispiel ist dies 4-2=2. Demnach ist die Position des gesuchten Strings <2,4>.

Die Anzahl der Übereinstimmungen (Matches) im Text mit dem Suchstring lässt sich anhand der Anzahl der Blätter im darunter liegenden Teilbaum ermitteln, die genau der Anzahl an Matches entspricht. „AS“ zum Beispiel kommt im String „ANASTASIA“ zweimal vor. Um die Position aller Matches zu berechnen muss die Länge aller Kanten ermittelt werden, was durch Abzug der jeweiligen Anfangsposition *a* von der jeweiligen Endposition *e* geschieht. Dann werden die Kantenlängen addiert, wodurch man die Gesamtkantenlänge erhält. Diese wird von der Gesamtlänge des Textes *l* abgezogen: $l-(e-a)_0+(e-a)_n$. Das Ergebnis ist die Anfangsposition des berechneten Matches. Im Beispiel sind das 10-(10-7)+(4-3)+(1-0)=5 und 10-(10-4)+(4-3)+(1-0)=2, woraus sich ergibt, dass die Matches an Position <2,4> und <5,7> zu finden sind.

3 Sprachtypen

Im Verlauf der Zeit gab es viele Versuche, die Sprachen in verschiedene Typen zu unterteilen. Vor allem im 19. Jahrhundert sind hierzu einige Aufsätze erschienen, wobei zwei Wissenschaftler bis heute von besonderer Bedeutung sind. August Schleicher hat mit seiner Dreiteilung den Grundstein für die heute verbreitetste Typologie gelegt. Er hat zwischen den einsilbigen, anleimenden und flektierenden Sprachen unterschieden, wobei die einsilbigen Sprachen heute als isolierend bezeichnet werden und die anleimenden als agglutinierend. (vgl. Hajdu:1987, S346) Humboldt hat diese Einteilung schließlich weiterentwickelt und um den inkorporierenden oder polysynthetischen Sprachtypus ergänzt.

Unter *isolierende Sprachen* werden Sprachen verstanden, deren Wörter keine, oder nur wenig Flexion aufweisen und grammatische Relationen daher auf anderem Wege, wie der Wortreihenfolge, Pausen oder dem Tonfall hergestellt werden müssen. Dies bringt mit sich, dass die Reihenfolge der Wörter in diesen Sprachen meistens sehr streng ist und nicht verändert werden kann, oder eine Veränderung einen deutlichen Bedeutungsunterschied mit sich bringt. Häufige Beispiele hierfür sind Vietnamesisch und die chinesischen Sprachen, aber auch das Englische weist einen tendenziell isolierenden Charakter auf:

(2) The man loves the woman. ↔ The woman loves the man.
Der Mann liebt die Frau. ↔ Die Frau liebt den Mann.

Flektierende Sprachen zeichnen sich dadurch aus, dass Lautwechsel im Stamm auftreten und Affixe mit dem Stamm verschmelzen können. Außerdem haben Affixe oft mehr als eine grammatische Funktion, wodurch komplexe Flexionssysteme vorherrschen in denen die einzelnen Funktionen nicht vollständig vom Stamm getrennt werden können. Das bietet jedoch im Gegensatz zu den isolierenden Sprachen einen deutlich größeren Spielraum in der Wortreihenfolge, wodurch diese für andere Inhalte, als der grammatischen Relation, wie etwa Betonung oder Satzmodus (Aussage, Frage, Ausruf) zur Verfügung steht. Zu den flektierenden Sprachen gehören beispielsweise Latein oder Polnisch, aber auch Deutsch hat einen deutlichen flektierenden Charakter.

(3) ich erfahre ↔ er erfährt ↔ ihr erfahrt
1.Pers.Sg ↔ 3.Pers.Sg ↔ 2.Pers.Pl

Bei *agglutinierenden Sprachen* dagegen gibt es für jede grammatische Funktion klar erkennbare Morpheme, die vollständig von anderen Morphemen abtrennbar sind und auch nur eine einzige Bedeutung oder Funktion enthalten. Variationen von verschiedenen Morphemen für eine Funktion können normalerweise nur durch phonologische Phänomene, wie Vokalharmonien oder phonetische Restriktionen entstehen. Auch hier ist die grammatische Relation nicht an die Wortreihenfolge gebunden. Meistgenannte agglutinierende Sprachen sind Finnisch und Türkisch.

(4) Türkisch:
köy-ler-in
Dorf-PL-GEN
der Dörfer (Steinbergs:1997, 381)

Die *inkorporierenden* oder *polysynthetischen Sprachen* könnten als Untergruppe der agglutinierenden Sprachen angesehen werden, haben jedoch die Besonderheit, dass auch semantische Bedeutungen in Wörter „inkorporiert“, also integriert werden können, wodurch sich teilweise ganze Sätze durch ein einzelnes Wort ausdrücken lassen. Wegen der daraus resultierenden speziellen Gegebenheiten werden sie daher in der Regel getrennt von den

agglutinierenden Sprachen betrachtet.

(5) Beaver:
náá-na-dy-u-s-'édz-esé
ITER-2sO-QUL-FUT-1sS-kick-FUT
I'm going to kick you (repeatedly). (Jung:2010, S6)

„Diese Einteilung der Sprachen in vier Typen trägt die Gefahr in sich, daß die isolierenden, agglutinierenden, polysynthetischen und flektierenden Sprachen als grundsätzlich voneinander abweichende, miteinander nicht in Übereinstimmung zu bringende Typen betrachtet werden.“ (Hajdu:1987 S351)

Es gibt keine bekannte Sprache, die keine Phänomene eines zweiten Sprachtypen enthält, auch wenn manche Sprachen, wie etwa das isolierende Vietnamesisch, einem „Idealtypen“ sehr nahe kommen. Ein gutes Beispiel für die Problematik bei der Einteilung ist die deutsche Sprache, das Phänomene aller vier Typen enthält.

So sind die Artikel zum Einen isoliert vom Substantiv, tragen aber gleichzeitig mehrere grammatische Funktionen (Genus, Numerus, Kasus). Dadurch lässt sich diese Wortart nicht eindeutig dem isolierenden oder dem flektierenden Sprachtypen zuordnen.

Das Verb ist in dieser Hinsicht besonders interessant, da hier gleich drei Sprachtypen nachweisbar sind. Zum Einen gibt es klar flektierende Strukturen, wie zum Beispiel bei „treffen – traf – getroffen“. Zum Anderen findet sich aber auch der agglutinierende Typ wieder, etwa bei dem Verb „hauen“:

(6) ich hau-e ↔ ich hau-t-e
du hau-st ↔ du hau-t-e-st
er hau-t ↔ er hau-t-e

Hierbei wird deutlich, dass das „t“ in der rechten Spalte nur die grammatische Funktion „Präteritum“ steht, während „e“ und „st“ für die jeweilige Person stehen. Bei der zweiten und dritten Person zeigen sich zwei phonologische Regeln, da in der deutschen Sprache nicht zwei „t“ hintereinander auftreten können, weshalb das „t“ für die Person durch ein „e“ ersetzt wird und ein „e“ in die zweite Person eingebaut wird, was aber nur, vereinfacht ausgedrückt, der leichteren Aussprache dient. Es ist also trotz der Veränderungen ein agglutinierendes Schema.

Der dritte Sprachtyp, der sich beim Verb des Deutschen zeigt, ist der inkorporierende Typ in Form der Partikelverben. In den meisten Fällen entsprechen die Verbpartikel Präpositionen, wie bei „auffahren“ oder anfahren. Jedoch können auch andere semantische Inhalte integriert werden, etwa bei „radfahren“, bei dem ein Substantiv zu einem Verbpartikel geworden ist.
Der inkorporierende Typ zeigt sich auch durch die im Deutschen besonders vielfältige Möglichkeit der Kompositionsbildung. So können zum Beispiel zwei oder mehr Substantive miteinander verbunden werden, wie bei „Taschentuch“ oder „Taschentuchpackung“, was sich entsprechend beliebig fortsetzen lässt. Aber auch mit Adjektiven wie „gelbgrün“ oder Verben wie „gefriertrocknen“ sind Kompositionen möglich. Die einzelnen Wortarten lassen sich auch beliebig miteinander kombinieren, zum Beispiel bei der Verb-Substantiv-Komposition „Auffahrunfall“. Auch die Kombination mit Phrasen ist möglich, beispielsweise bei der „Augen-zu-und-durch-Politik“ (Donalies:2005, S72).
Mit anderen Worten, es zeigt sich, dass der zugeordnete Sprachtyp einer Sprache lediglich den dominierenden Typen wiedergibt und deshalb nur bedingt zur Unterscheidung unterschiedlicher Sprachen dienen kann. Es sollte jedoch festgehalten werden, dass es wahrscheinlicher ist, dass eine Sprache eine starre Wortstellungsstruktur aufweist, je stärker der isolierende Charakter ist.

3.1 Sprachen mit freier Wortstellung

Es gibt einige Sprachen, die als Sprachen mit freier Wortstellung bezeichnet werden, doch bei näherer Betrachtung ergeben sich beträchtliche Unterschiede in der Freiheit der Wortstellung auch bei diesen Sprachen. So ist etwa die Wortstellung im Polnischen grundsätzlich frei was die Syntax betrifft, kann jedoch in bestimmten Fällen dennoch syntaktische Bedeutung erhalten, nämlich, wenn die Zuordnung anders nicht eindeutig ist (was allerdings nur selten vorkommt):

(7) Kinder mögen Tiere. Dzieci lubią zwierzęta.
Tiere mögen Kinder. Zwierzęta lubią dzieci. (Engel 1999: 492)

Außerdem kann die Nominalphrase nicht auseinander gebrochen werden. Adjektiv und Substantiv können also nicht voneinander getrennt werden (innerhalb der Phrase aber in der Reihenfolge frei variiert). Außerdem steht das Wort '*nie'* (= „nicht“) immer vor dem negierten

Wort. Wird der Satz negiert, so orientiert sich das Wort am Verb. Und noch etwas fällt bei der Wortstellung im Polnischen auf; Die Wortreihenfolge hat nämlich durchaus eine Bedeutung. Sie wird benutzt, um Topik und Fokus zu kennzeichnen.

Das ***Topik*** bezeichnet dabei den Referenten eines Satzes, also das, worum es in diesem Satz geht, im Gegensatz zum Comment, also das was über den Referenten ausgesagt wird. Dieser kann auch im Deutschen durch die Umstellung des Satzes gekennzeichnet werden:

(8) Einen Ball habe ich gefunden.

In einigen Sprachen, wie Japanisch, wird das Topik auch per Topikmarker markiert. Zu beachten in Beispiel (9) ist, dass der Topikmarker in der japanischen Sprache postpositional ist, also dem topikalisierten Element folgt.

(9)	haha	wa	ko-no	hon	o	kat-te	kure-ta
	Mutter	TOP	dies-ADJ	Buch	AKK	kauf-GER	geb-PRT

„Mutter hat mir dieses Buch gekauft." (Lehmann:2010 3.3.4)

Der ***Fokus*** liegt dort, wo Präsupposition von Assertion getrennt wird. In den meisten germanischen Sprachen, wie Deutsch oder Englisch, wird dies durch Akzentuierung ausgedrückt. (vgl. Lehmann:2010 3.3.5) Präsupposition ist das vorausgesetzte Weltwissen, während Assertion die neue Information bezeichnet. (vgl. Lehmann:2010 3.3.2)

Andere Sprachen sind dagegen deutlich freier, wie etwa das Nez Percé, in dem die Nominalphrase durchbrochen werden kann, wie im Beispiel (8), wo das Substantiv am Ende und das dazugehörende Adjektiv am Anfang des Satzes steht.

(10)	yu'ús-ne	taxc	ki-nm	ta'c	'iyéext	'a-anyáa-'n-yo'	na'-tóota-p
	poor-OBJ	soon	this-GEN	good	broth	3SAP-make-BEN-FUT	my-**father**-OBJ

'Soon I will make of this a good broth for my **poor father**' (Rude 1992: 196)[4]

Es fällt auf, dass es auf der einen Seite Sprachen gibt, die zwar wie Polnisch eine Grundwortstellung haben, diese aber häufig durchbrechen, während bei anderen Sprachen, wie eben auch das oben erwähnte Nez Percé keine Wortstellung im Speziellen präferiert wird.

4 Das Beispiel stammt ursprünglich aus Phinney (1934), wo sich die Rechtschreibung jedoch etwas unterscheidet: „yʊ'u'sna taxts kınm ta'ts 'ıya'xt 'a-anya'nyo' na'-to'tap." (S. 268)

„In a number of languages, the order of constituents does not reflect their syntactic functions at all, but rather their pragmatic functions“ (Mithun 1992: 58). Daher können die Sprachen nach Mithun in syntaktisch basiert („syntactically based“) und pragmatisch basiert („pragmatically based“) eingeteilt werden. „[In] syntactically based languages, pragmatic reordering is highly marked [for] unusual situation[s] [and] pragmatic reordering is usually assumed to result in a theme-rheme order“[5]. „In pragmatically based languages, on the other hand, all ordering reflects pragmatic considerations. [These] languages are typically highly polysynthetic.“ (Mithun 1992: 58f) Dass pragmatisch basierte Sprachen meistens polysynthetische Sprachen sind, erklärt Mithun damit, dass diese Sprachen über komplexe gebundene Pronomen verfügen, die die primären grammatischen Relationen in Bezug auf das Verb tragen.

Eine andere, deutlich detailliertere Einteilung nimmt Koktová (1996) vor, indem sie eine Tiefen- und eine Oberflächenwortstruktur („deep“ / „surface“) annimmt, die beide fest („fixed“) oder frei („free“) sein können. Die Tiefenwortstruktur bezieht sich auf die Ebene von Topik und Fokus, während die Oberflächenwortstruktur auf die Wortarten Bezug nimmt. Daraus leitet Koktová fünf Wortstellungstypen ab:

I. Die ***feste Tiefenwortstruktur*** („deep fixed word order“) bezieht sich auf eine Festlegung in der Reihenfolge von Topik und Fokus. „Here belong the hypothetical ordering of the five communicative-information parts of the sentence, from Noncontrastive Topic as the best known and best accessible part, through Contrastive Topic to Focus as the less known part.“ (Koktová 1996: 412) Koktová zählt also fünf kommunikative Informationen auf, von denen sie allerdings nur drei nennt, nämlich den nichtkontrastiven Topik, den kontrastiven Topik und den Fokus. Da eine klare Trennung zwischen Topik und Fokus allerdings nicht immer gegeben ist, ist ihre Einteilung problematisch. Dennoch lohnt es sich, sich hiermit näher zu beschäftigen, da die Unterteilung Einfluss auf die Wortstruktur haben kann. Wahrscheinlich ist die Unterteilung der Einheiten in der Tiefenwortstruktur deshalb so schwer, weil es je nach Sprache unterschiedliche „Sichtweisen“ geben kann.

II. Die ***freie Tiefenwortstruktur*** („deep free word order“) ist sozusagen das Komplement zur

5 Thema und Rhema sind neben Fokus und Topik eine weitere Möglichkeit der Einteilung der Informationsstruktur von Sätzen. Da diese Begriffe jedoch auf zwei verschiedene Arten interpretiert werden können und diese auch in der Literatur nicht immer klar voneinander abgegrenzt werden, verzichte ich auf ihren Gebrauch. (vgl. Bußmann 2002: 695f)

festen Tiefenwortstruktur und bezeichnet die freie Variation von Elementen der Tiefenwortstruktur. Als Beispiel kann hier in den meisten Sprachen die freie Reihenfolge der adverbialen Bestimmungen von Zeit und Ort gesehen werden (vgl. Koktová 1996: 414).

III. Die ***feste Oberflächenwortstruktur*** („surface fixed word order“) bezieht sich auf die Festlegung der Wortarten auf bestimmte Bereiche. Dies kann die allgemeine Wortstruktur in einem Satz betreffen, die in einigen Sprachen immer S-V-O[6] (z. B. Englisch, vgl. Koktová 1996: 414), in anderen Sprachen aber beispielsweise immer S-O-V ist. Aber auch die Satzklammern im Deutschen sind ein Fall fester Oberflächenwortstruktur: „A prominent case of surface fixed word order in German is the frame construction, where certain verbal forms [...] occur after the nonverbal part of the Focus“.

IVa. Die ***freie Oberflächenwortstruktur mit Bezug auf die Tiefenwortstruktur*** („surface free word order corresponding to deep word order“) bezieht sich auf die Durchbrechung („relaxation“) einer festen Oberflächenwortstruktur und einer Reorganisation von ihr durch die Tiefenwortstruktur. „[The] relaxations of fixed word order are more frequent [...] if the ‚moved‘ expressions [...] are pragmatically (communicatively) relevant enough to be ‚moved‘ from their fixed positions to the positions which roughly correspond to their deep positions.“ (Koktová 1996: 416)

IVb. Die ***freie Oberflächenwortstruktur ohne Bezug auf die Tiefenwortstruktur*** („surface free word order not corresponding to deep word order“) setzt im Gegensatz zum Typ IVa. keine festen Oberflächenwortstrukturen voraus, sondern bezieht sich vielmehr auf das Verschieben von Elementen im Satz, um die Prominenz hervorzuheben, auch bzw. gerade wenn dies nicht durch die Tiefenwortstruktur initiiert wird. „Here belong various shifts of communicatively important elements to communicatively prominent positions, such as the sentence-initial or the preverbal position.“ (Koktová 1996: 416) Dieser Wortstellungstyp ist besonders problematisch, da hier keine klaren Regeln bestehen, wie und wann entsprechende Phrasen verschoben oder sogar auseinander gerissen werden wie etwa beim Nez Percé in Beispiel (10).

6 Die Buchstaben S, O und V bezeichnen das Subjekt (S), das Objekt (O) und das Verb (V). Es ist zwar eine gemischte Notation, da Subjekt und Objekt Satzteile bezeichen, das Verb jedoch eine Wortart ist, sie ist jedoch in der Typologie üblich.

Die Einteilung der Wortstellungstypen stellt kein Versuch dar, die Sprachen einzuteilen, wie dies etwa bei den Sprachtypen der Fall ist. Eher werden verschiedene Strategien beschrieben, die Wortstellung zur Informationsübermittlung zu gebrauchen. „The phenomena in question are only a matter of degree, with respect to the synergetic (self-regulating) and functional (weakly teleological) character of natural languages.“ (Koktová 1996: 417)

Um zu klären, wie Sprachen mit freier Wortstellung funktionieren und was sich genau ändert, muss der Frage nachgegangen werden was einen Sprecher überhaupt dazu bringt, die Wortstellung umzustellen. Dieser Frage ist Payne (1992b) nachgegangen. Sie hat vier Gründe herausgearbeitet, ob diese vollständig sind, bleibt zu überprüfen. Diese sind im Einzelnen

(1) dass der Hörer die übermittelte Information so versteht, wie der Sprecher wünscht, dass sie verstanden wird, also zur Verdeutlichung oder Modifikation der Bedeutung einer Aussage,
(2) dass der Sprecher vermutet, dass etwas genau konträr zum Wissen des Hörers ist, also dass der Sprecher wünscht, das fehlerhafte Wissen des Hörers zu korrigieren,
(3) dass ein Satz einen Diskurs beendet,
(4) dass ein Teil der Bedeutung einer Aussage die Information beinhaltet, dass die Information relativ zur Information in einer anderen Aussage ist.

Es handelt sich also um Gründe, die nicht direkt die eigentliche Aussage (Proposition) betreffen, sondern immer den Wunsch des Sprechers auf eine bestimmte Reaktion des Hörers. Sprechakttheoretisch ausgedrückt findet mithilfe der Wortumstellung eine Modifikation des illokutionären Aktes statt, während die Proposition unverändert bleibt. Damit spielt sie eine besondere Bedeutung bei der Interpretation eines Textes und damit auch für die Anwendungsgebiete der Suffix Trees.

Die Einteilung von Koktová hilft dabei die problematischen Bereiche einer Sprache abzugrenzen und möglicherweise die richtigen Lösungsstrategien dazu zu entwickeln, während Mithun eine grundsätzliche Einteilung vornimmt, bei der ersichtlich wird, dass wahrscheinlich unterschiedliche Strategien bei syntaktischen und pragmatischen Sprachen angewendet werden muss. Wahrscheinlich ist dies jedoch ein Kontinuum bei der eine klare Einteilung nicht immer möglich ist, weshalb es trotzdem sinnvoll ist, sich über eine Gesamtlösung Gedanken zu machen. Allerdings dürften Teilstrategien für einzelne Sprachen effizienter sein, als eine entsprechende Generalstrategie.

4 Suffix Trees und Sprachen mit flexibler Wortstellung

4.1 Probleme

Wie im vorhergehenden Abschnitt dargestellt, gibt es Sprachen, die sehr flexibel mit den Wörtern innerhalb eines Satzes umgehen (können). Will man nun eine solche Sprache mit Suffix Trees verarbeiten, stößt man auf das Problem, dass bei gleicher Bedeutung unterschiedliche Bäume entstehen, was die Auswertung und Suche, vor allem für Übersetzungen und beim Information Retrieval deutlich erschwert.
Zusätzlich kann Zusammengehöriges auseinander liegen. Das bedeutet, dass Phrasen und Wörter voneinander getrennt sein können und im Suffix Tree daher nicht als zusammengehörig erkannt werden. Fehlerhafte Zuordnungen sind dann die Folge.

4.2 Lösungsansätze

In jedem Fall ist eine Vorverarbeitung nötig, da die Suffix Trees selbst nur eine Datenstruktur darstellen. Wie diese Vorverarbeitung allerdings aussieht, unterscheidet sich hinsichtlich der ausgewählten Lösungsstrategie. Alle von mir erarbeiteten Ansätze enthalten noch ungelöste Probleme und sind deshalb nicht als endgültig zu sehen. Weitere Überlegungen und eine praktische Erprobung sind also erforderlich, um eine praktikable Lösung zu finden.

Die Lösungsansätze sind im Einzelnen

a) der **statistische Ansatz**, bei dem die häufigste Wortreihenfolge betrachtet wird,
b) der **Reihenfolge verändernde Ansatz**, bei dem die Wortreihenfolge entsprechend einer vorher bestimmten Reihenfolge und mithilfe der Flexive geändert wird und
c) der **Token[7] verknüpfende Ansatz**, bei dem die einzelnen Tokens ohne feste Reihenfolge miteinander verknüpft werden.

a) Statistischer Ansatz

Da auch Sprachen mit einer flexiblen Wortstellung gewissen Restriktionen unterliegen, lassen sich Abweichungen in vielen Sprachen erkennen. Bleiben die Sätze mit „Sonderwortstellung“ unberücksichtigt, sind damit die daraus resultierenden Probleme ausgeschaltet. Die Sätze in

7 *Tokens* sind alle gegebenen Zeichenketten, die in einem Satz durch Leerstellen oder Satzzeichen voneinander getrennt sind.

einem Musterkorpus müssen dafür zunächst geparst und Wortstrukturmuster erstellt werden. Diese Wortstrukturmuster müssen dann miteinander verglichen werden. Gibt es mehrere mögliche Wortstrukturmuster muss zusätzlich überprüft werden, ob bestimmte Bedingungen (z. B. die Satzarten im Deutschen) unterschiedliche Muster erzeugen. In einem zu verarbeitenden Korpus werden die Sätze mit den entsprechenden Mustern verglichen und die abweichenden Sätze aussortiert. Dies funktioniert immer dann gut, wenn eine feste Tiefenwortstruktur existiert, in allen anderen Fällen fällt dieser Ansatz allerdings weg, da eine Grundstruktur vorhanden sein muss, damit eine Statistik überhaupt aufgestellt werden kann. Auch bei in bestimmten Fällen auseinandergezogenen Phrasen oder Wörtern, wie den deutschen Partikelverben hilft der statistische Ansatz nicht weiter. Der Nachteil dieses Ansatzes besteht darin, dass Informationen verloren gehen, die in den unberücksichtigten Sätzen stehen. Bei Information-Retrieval-Anwendungen können so wichtige Informationen unsichtbar bleiben und Übersetzungen bleiben so unvollständig.

b) Reihenfolge verändernder Ansatz

Entweder hat eine Sprache eine strenge Wortreihenfolge, oder die Wörter haben Flektive, die die Zugehörigkeit zu einer bestimmten Phrase anzeigen. Im Vorfeld muss hier ein Musterkorpus wie in Ansatz a) geparst werden. In einem zweiten Schritt werden dann die Flektive ermittelt, die zusammen gehören und ein entsprechendes Lexikon erstellt, das die Zugehörigkeiten enthält. Im zu verarbeitenden Korpus können die entsprechenden Wörter, die zwar zu einer Phrase gehören, aber nicht beieinander stehen anhand des angelegten Lexikons zusammen geschoben werden. So muss nicht auf die Sätze verzichtet werden, alle werden erfasst und auch pragmatische Sprachen können, anders als im statistischen Ansatz, verarbeitet werden, da ein Wortstrukturmuster dabei nicht nötig ist. Problematisch ist bei diesem Ansatz, dass der Text manipuliert wird und daher die Illokution des Verfassers nicht mehr berücksichtigt werden kann. Dies kann vor allem bei Übersetzungen zu Fehlern führen, da hier die Illokution durchaus von Bedeutung ist. Allerdings bleiben alle Informationen erhalten und durch die Zusammenführung der Phrasen werden Fehler in der Proposition vermieden.

c) Tokens verknüpfender Ansatz

Pragmatische Sprachen lassen sich mit dem statistischen Ansatz nicht umsetzen, da kein bevorzugtes Wortstellungsmuster existiert und sie somit nicht statistisch ermittelbar ist. Mit dem zweiten Ansatz sind pragmatische Sprachen zwar grundsätzlich verarbeitbar, er hat jedoch das Problem, dass die Illokution unberücksichtigt bleibt, diese aber gerade in pragmatischen Sprachen eine stärkere Rolle spielt. Deshalb ist auch Ansatz b) nicht geeignet um solche Sprachen zu verarbeiten. Eine weitere Möglichkeit ist daher, die einzelnen Tokens unabhängig voneinander in unterschiedlichen Suffix Trees zu speichern und diese Bäume dann auf zwei Ebenen miteinander zu verknüpfen, nämlich auf der Wortstellungsebene und einmal auf der syntaktischen Ebene. Dadurch werden sowohl die Wortreihenfolge, als auch die Zugehörigkeiten innerhalb der Phrasen berücksichtigt. Es müssen also nach dem Erstellen der Suffix Trees zwei zusätzliche Speicherstrukturen eingefügt werden. Diese könnten möglicherweise als Listen realisiert werden. Das Problem bei diesem Ansatz ist, dass der Einsatz der Suffix Trees dadurch erheblich eingeschränkt ist und wahrscheinlich verschachtelte Algorithmen mit erheblichem Zeitaufwand nötig sind, was die Vorteile der Suffix Trees deutlich schmälert, wenn nicht sogar aufhebt.

5 Ausblick

Alle drei Ansätze sind nicht ideal für den Einsatz in der Praxis, da sie alle Mängel aufweisen, die leider noch nicht beseitigt werden konnten. Der statistische Ansatz eignet sich nur für syntaktische Sprachen und auch dort nur in Bereichen mit fester Tiefenwortstruktur. Auch der Reihenfolge verändernde Ansatz ist nur bedingt für pragmatische Sprachen geeignet. Der Tokens verknüpfende Ansatz dagegen müsste auf seine grundsätzliche Praxistauglichkeit getestet werden.

Es sollten also noch tiefergehende Überlegungen zu den Strategien und zu möglichen Kombinationen erarbeitet werden. Dazu gehört auch die Überlegung, wie diese praktisch umgesetzt werden können. Hinterher müssen diese schließlich realisiert und in ihrer Effizienz (u. a. Zeitaufwand und Speicherbedarf) geprüft und verglichen werden.

Bibliographie

BUSSMANN, Hadumod (Hrsg.). 2002. Lexikon der Sprachwissenschaft. Stuttgard: 3. aktual. u. erw. Aufl., Kröner.

DONALIES, Elke. 2005. *Die Wortbildung des Deutschen – Ein Überblick*. (Studien zur Deutschen Sprache, 27.) Tübingen: Narr.

ENGEL, Ulrich, et. al. 1999. *Deutsch-polnische kontrastive Grammatik*. Heidelberg: Groos.

HAJDU, Péter, DOMOKOS, Péter. 1987. *Die uralischen Sprachen und Literaturen*. Budapest: Buske.

HUMBOLDT, Wilhelm von. 1836. *Über die Kawi-Sprache auf der Insel Java.* Berlin: Königliche Akademie der Wissenschaften.

JÄGER, Gerhard. 2006. *Morphologische Sprachklassifikation.* http://www2.sfs.uni-tuebingen.de/jaeger/lehre/ws0607/sprachenDerWelt/morphologischeTypologie.ppt 30. Dez. 2009.

JAKOB, MELANIE. 2010. *Suffix Trees: Simple Algorithm and Applications.* http://www.bio.ifi.lmu.de/webfm_send/2083 30. Aug. 2010.

JUNG, Dagmar. 2010. „Layers of Classification within the Athabaskan verb." (Handout). unveröffentlicht.

KAY, Martin. 2003. *Substring alignment using Suffix Trees.* http://www.stanford.edu/~mjkay/CYCLING.pdf 30. Dez. 2009.

KAY, Martin. 2004 *Aligning Sentence Parts.* http://www.stanford.edu/class/linguist139p/ICON.pdf 30. Dez. 2009.

KOKTOVÁ, Eva. „Word order and typology." In: PALEK, Bohumil (Hrsg.). 1996. Typology: prototypes, item orderings and universals. (Acta Universitatis Carolinae: Philologica, 1996,3/4). Prag.

LEHMANN, Christian. 2010. *Morphologie und Syntax.* http://www.christianlehmann.eu/ling/lg_system/grammar/morph_syn/index.html 30. Aug. 2010.

LÓPEZ, Justo Fernández. 2009. *Sprachtypologie*. http://culturitalia.uibk.ac.at/hispanoteca/Lexikon%20der%20Linguistik/sp/SPRACHTYPOLOGIE%20%20%20Tipolog%C3%ADa%20de%20las%20lenguas.htm 30. Dez. 2009

MITHUN, Marianne. „Is Basic Word Order Universal?“ In: PAYNE, Doris L. (Hrsg.). 1992a. 15-61

NELSON, Mark. 1996. *Fast String Searching With Suffix Trees*. http://marknelson.us/1996/08/01/suffix-trees/ 30. Dez. 2009

O'GRADY, William, et al (Hrsg.). 1997. *Contemporary Linguistics*. An Introduction. London (u. a.): Longman.

PAYNE, Doris L. (Hrsg.). 1992a. Pragmatics of word order flexibility. (Typological studies in language, 22). Amsterdam (u. a.): Benjamins.

PAYNE, Doris L. 1992b „Introduction“ In: PAYNE, Doris L. (Hrsg.). 1992a. 1-13.

PHINNEY, Archie. 1934. Nez Percé Texts. (Columbia University Constributions to Anthropology, 25). New York: Columbia University Press.

RUDE, Noel. „Word Order and Topicality in Nez Perce“ In: PAYNE, Doris L. (Hrsg.). 1992a. 193-208.

STEINBERGS, Alexandra. „The classification of language.“ In: O'GRADY, William, et al (Hrsg.). 1997. 372-415.

UKKONEN, Esko. 1995. „On-line construction of suffix trees.“ *Algorithmica*. 14:249-260.

WERNER, Agnes. 2003. *Die Wortstellung in polnischen Deklarativsätzen mit transitiven und ditransitiven Verben und ihre Bedeutung für die Informationsstruktur*. http://www.uni-potsdam.de/u/slavistik/wsw/seminararbeiten/Werner_Informationsstruktur.pdf 30. Dez. 2009

WIKIPEDIA. IR. „Information Retrieval.“ 30. Dez. 2009. HTTP://DE.WIKIPEDIA.ORG/W/INDEX.PHP?TITLE=INFORMATION_RETRIEVAL&OLDID=84390634

WIKIPEDIA. LS. „Landau-Symbole.“ 10. Aug. 2010. http://de.wikipedia.org/w/index.php?title=Landau-Symbole&oldid=77677633

WIKIPEDIA. PT. „Patricia-Trie.“ 28. Jul. 2010. http://de.wikipedia.org/w/index.php?title=Patricia-Trie&oldid=77161936

WIKIPEDIA. Sb. „Suffixbaum.“ 30. Dez. 2009. http://de.wikipedia.org/w/index.php?title=Suffixbaum&oldid=68604491

WIKIPEDIA. T. „Trie.“ 25. Apr. 2010. http://de.wikipedia.org/w/index.php?title=Trie&oldid=73587048