Extraktion und Verifikation von Subkategorisierungsmustern für französische Verben

Tanja GAUSTAD, Groningen, The Netherlands

Zusammenfassung
Die halbautomatische Extraktion und Verifikation von Subkategorisierungsmustern bringt viele Vorteile mit sich bei Erstellung von syntaktischen Wörterbüchern. Im hier diskutierten Ansatz werden reguläre Suchroutinen mit 'einfachen' linguistischen Verfahren, d.h Verfahren, die mit wenig linguistischem Wissen arbeiten, kombiniert, um Belege aus annotierten Corpora zu entrahieren. Die Belege werden anschliessend manuell gesichtet und zu einem Lexikoneintrag abstrahiert. Probleme treten vor allem im Zusammenhang mit der Erkennung von Mehrwortausdrücken auf.

1 Einführung

Im folgenden Beitrag geht es um die halbautomatische Extraktion von Subkategorisierungsinformationen für die Bildung von syntaktischen Wörterbüchern französischer Verben[1]. Subkategorisierungsinformationen zu Verben leisten einen essentiellen Beitrag zur detaillierten Beschreibung der in Wörterbüchern zu erfassenden Lexeme. Sie geben Aufschluss über die Art und Zahl der Komplemente, über deren Position (Distribution) sowie über mögliche Kollokationen (Subjekt–Verb, Verb–Objekt)[2]. Dazu wird ein corpusbasierter Ansatz gewählt, der es ermöglicht, von den Corpustexten über Belegsammlungen zu Lexikoneinträgen zu abstrahieren. Ein corpusbasierter Ansatz bringt den Vorteil, dass Belegmaterial nach zuvor festgelegten Kriterien/Suchmustern extrahiert und anschliessend, nach manueller Überprüfung durch den Linguisten oder Lexikographen, in das jeweilige Wörterbuch aufgenommen werden kann.

2 'Einfache' Verfahren

Die Motivation zu einem corpusbasierten Ansatz geht aus der Idee hervor, dass durch die Nutzung verhältnismässig 'einfacher' Verfahren auf kleinen, syntaktisch analysierten Textmengen relativ gute Erfolge erzielt werden können. Der Einsatz 'einfacher' Verfahren, d.h. Verfahren, die keine höheren linguistischen Informationen voraussetzen, erklärt sich durch die Tatsache, dass viele robuste Grammatiken oft genau die gesuchten Informationen (in diesem Fall syntaktische Subkategorisierungsmuster) implizit voraussetzen. In der hier untersuchten Vorgehensweise—mit elementaren Annotationen möglichst viele, korrekte Belege aus Corpora zu extrahieren—kann die Akquisition des Lexikonmaterials über den Einsatz zweier Hilfsmittel vonstatten gehen: über statistische Verfahren (z.B. stochastische Grammatiken), die ohne oder mit sehr wenig linguistischem Wissen arbeiten, oder über reguläre Grammatiken, wie z.B. die von [Gross, 1986] oder [Silberztein, 1993] für verschiedene Aufgaben der Textprozessierung eingesetzten 'lokalen Grammatiken'. Im vorliegenden Beitrag werden die Möglichkeiten und Probleme, welche die Arbeit mit 'lokalen' Grammatiken in Bezug auf Französische Verben birgt, illustriert und ausgewertet.

	CQP-Ausdrücke	Beispiel	Kommentare
1	`[word=''me|m'|te|t'|nous|vous|lui|leur'' & pos=''PRO:pers:clit|PRO:pers:refl'']`	*leur*	pron. indir. Objekt
2	`[lemma=''avoir'' & pos=''VER.*'']?`	*a*	Hilfsverb (opt.)
3	`[pos=''ADV.*'']{0,2}`	*jamais*	Adverb (opt.)
4	`[lemma=''apprendre'' & pos=''VER.*'']`	*appris*	Verb

Tabelle 1: Beispiel einer Suchroutine

3 Prämissen

Als generelle Prämissen gelten: die verwendeten Corpora sind linguistisch vorverarbeitet und annotiert auf Wort- und Satzgrenzen (*Tokens*), auf Wortklassen (*Tagging*) sowie auf Lemmata. Diese bestehenden Informationen werden bei der Verwendung regulärer Grammatiken und der Entwicklung zugehöriger Suchmuster eingesetzt[3]. Die entwickelten Suchroutinen beziehen ausserdem die Oberflächenreihenfolge der zu analysierenden Wortsequenz, optionale Teilsequenzen innerhalb der gesuchten Wortsequenz sowie die Kombination von Wortform-, Lemma- und Kategorieangaben oder deren expliziten Ausschluss mit ein. Tabelle 1 zeigt ein Beispiel einer Suchroutine.

Wichtig anzumerken ist dabei, dass die regulären Suchmuster in ihrer Ausdruckskraft beschränkt sind, da sie lediglich auf morphosyntaktische Merkmale, nicht aber auf grammatische Funktionen zurückgreifen können. Der Gebrauch einer KWIC-Suchmaschine (*Key Word In Context*), um Belege aus den Corpora zu extrahieren, bietet ausserdem keine Möglichkeit, nach allgemeinen syntaktischen Rahmen zu suchen. Es muss explizit nach einem bestimmten Lemma in einem spezifischen syntaktischen Rahmen gesucht werden.

4 Arten der Extraktion

Die Extraktion von Subkategorisierungen nimmt zwei Formen an. Entweder wird von der Annahme ausgegangen, dass keinerlei linguistische Informationen zu den gesuchten Konstruktionen bestehen, was bedeutet, dass sämtliche Subkategorisierungsdaten von Grund auf gelernt werden müssen (siehe [Eckle-Kohler, 1998] für Deutsch). Oder aber die Subkategorisierungsmuster werden mit Hilfe bestehender Syntaxwörterbücher bestimmt und anschliessend im Corpus verifiziert, was der im folgenden beschriebenen Vorgehensweise entspricht. Als allgemeines Suchprinzip gilt es bei beiden Ansätzen, Mengen von Kontexten zu suchen, die syntaktische Eigenschaften von Lexemen eindeutig illustrieren. Die Konsequenz daraus ist, dass die so erlangten Informationen nach syntaktischen Subkategorisierungstypen aufgebaut sind (und nicht semiasologisch).

Zur Verifikation von Subkatergorisierungsmustern benötigt man in einem ersten Schritt Angaben zu den möglichen Valenzformen eines Lexems, in diesem bestimmten Fall eines Verbs. Dabei leisten bestehende Syntaxlexika, wie beispielsweise dasjenige von [Busse and Dubost, 1989], viel Vorarbeit. Nach der Zusammenstellung der verschiedenen (erwarteten) Subkategorisierungstypen, werden diese in den Corpora verifiziert. Im Französischen

spielt dabei die Wortstellung eine zentrale Rolle, da nur (gewisse) pronominalisierte Formen kasusmarkiert sind und die grammatikalische Funktion einer Sequenz typischerweise von deren Position relativ zum (Haupt)Verb abhängt. Dies hat zur Folge, dass die entwickelten Suchmuster bei der Extraktion von Okkurenzen stark mit wortstellungsspezifischen Angaben arbeiten.

5 Wortstellungshomogene Corpora

Um jedoch überhaupt auf wortstellungsbezogene Muster zurückgreifen zu können, müssen vor der Extraktion der gesuchten Valenzkonstruktionen die verwendeten Corpora dahingehend verändert werden, dass sie wortstellungshomogen sind: alle vom Aussagesatz abweichenden Wort- und Konstituentenreihenfolgen sowie valenzverändernde Prozesse (z.B. Passiv) müssen ausgeschlossen werden (als Beispiel für das Englische siehe [Gahl, 1998]). Erst auf Basis eines solchen wortstellungshomogenen (Sub)Corpus ist es möglich, mit dem gewählten Ansatz und den zur Verfügung stehenden Mitteln Belegmaterial zu extrahieren. Im konkret untersuchten Fall französischer Verben handelt es sich bei den von der Extraktion ausgeschlossenen, wortstellunsgverändernden Sätzen um Passivkonstruktionen (*que deux langues soient apprises*), um gewisse idiomatische Wendungen (*apprendre par cœur*) und um sogenannte 'constructions contexte gauche', die Inversion (*que ne lui ai-je appris*), Relativsätze (*chose que le Président ne vous a pas apprise*) sowie Konstituentenfragen (*à qui est-ce qu'ils l'ont appris*) zusammenfassen.

6 Extraktion von Belegen

Bei der Extraktion der zu klassifizierenden Belege müssen auf der Ebene der Suchroutinen-Entwicklung folgende Punkte berücksichtigt werden. Die Reihenfolge der Komplemente ist im Zusammenhang der Oberflächenlinearität sehr wichtig: im Französischen können Infinitiv-Komplement und indirektes Objekt permutieren. Um beide Wortstellungsvarianten abzudecken, müssen zwei Suchroutinen eingesetzt werden[4]. Die folgenden Beispielsätze illustrieren die obgenannten Wortstellungsvarianten (Beispiel (1) und (2) die Abfolge Verb–direktes Objekt–Infinitiv-Komplement, Beispiel (3) die Abfolge Verb–Infinitiv-Komplement–direktes Objekt).

(1) *Cependant on néglige d'apprendre aux civils, aux spectateurs innocents à se protéger.*

(2) *. . . ce qui consiste à apprendre à des centaines de milliers de jeunes à devenir chômeurs.*

(3) *. . . tel Ryoji Takahasi, qui apprend à voler aux grues blanches du Japon.*

Nominale Komplemente lassen sich nicht über Kasus identifizieren, sondern lediglich über morphosyntaktische Merkmale wie Konjunktionen (*que, si*), Pronomina (*quand, où*) und Infinitiv-Partikel (*à, de*) sowie über Wortstellung (siehe Suchroutine in Tabelle 2 und Beispiele (1)–(3)).

Nur bei pronominalisierten Objekten ist eine Identifikation über den Kasus möglich (*le* vs. *lui*), wozu wiederum spezielle Suchmuster eingesetzt werden (wie z.B. in Tabelle 1). Hier noch einige Beispiele zur Illustration:

	CQP-Ausdrücke	Beispiel	Kommentare
1	`[lemma=''apprendre'' & pos=''VER.*'']`	*apprendre*	Verb
2	`[pos!=''PRE:1st|CON.*|PON.*]{0,3}`		keine komplexen Präpositionen
3	`[word=''\'{a}|au|aux''] [pos!=''VER.*'']{0,3} [pos=''N.*|ADJ.*|PRO.*'']`	*à ses joueurs*	Indirektes Objekt
4	`[pos!=''CON:coo|PON.*'']`		Keine Koordination
5	`[word=''\'{a}'']`	*à*	Infinitiv-Partikel
6	`[pos=''ADV.*'']{0,2}`	*mieux*	Adverb (opt.)
7	`[pos=''PRO.*'' |word=''rien|beaucoup'']?`		Pron. dir. Objekt od. *rien/beaucoup*
8	`[pos=''VER:infi'']`	*respecter*	Infinitiv

Tabelle 2: Suchroutine für nominale Komplemente

(4) *Très tôt dans la vie, on nous apprend, entre autres choses, à dire au moins deux mots: "oui" et "non".*

(5) *Simplement parce qu'on ne leur a jamais appris à seulement exprimer leurs sentiments.*

Die Suchroutinen selbst werden durch die Aufteilung des wortstellungshomogenen (Sub)Corpus in weitere (Sub)Corpora strukturiert. Die dabei gewählte binäre Vorgehensweise ist nicht linguistisch motiviert, sondern erleichtert die Komplement-Bildung in einem prozessorientierten Umfeld. In einem ersten Schritt werden Suchroutinen, die finite oder infinite Komplemente extrahieren[5], von denjenigen, die nominale Komplemente suchen, unterschieden. Als zweites strukturierendes Kriterium wird die Valenz eingesetzt (2-stellig vs. 3-stellig). Des weiteren werden syntaktisch voll realisierte nominale von pronominalen Komplementen unterschieden und in einem letzten Unterteilungsschritt werden Wortstellungsvarianten separiert.

7 Resultate

Zusammenfassend lässt sich sagen, dass alle durch die Syntaxlexika bestimmten und gesuchten Konstruktionen in den analysierten Corpora belegt sind (siehe Tabelle 3). Die statistischen Auswertungen geben auch Aufschluss über die Verteilung der extrahierten Konstruktionsmuster sowie über die Gewichtung von Reihenfolge-Variationen bei variablen Komplementen (z.B. Verb–Infintiv-Komplement–indirektes Objekt vs. Verb–indirektes Objekt–Infinitiv).

Der Rest, der mit diesem Verfahren nicht klassifizierten Okkurenzen beläuft sich auf (lediglich) 15-20%, wobei sich dieser hauptsächlich auf falsch getaggte Wörter sowie die nun zu diskutierenden 'false positives' zurückführen lässt.

apprendre	Valenz	9831
+ Infinitiv	2	214
+ Infinitiv + indirektes Objekt	3	156
+ Infinitiv + pron. indirektes Objekt	3	81
+ *que*-Komplement	2	4130
+ *que*-Komplement + indirektes Objekt	3	26
+ *que*-Komplement + pron. indirektes Objekt	3	748
+ direktes Objekt + indirektes Objekt	3	271
+ direktes Objekt + pron. indirektes Objekt	3	357
+ direktes Objekt	2	1350
+ pron. direktes Objekt	2	333
Unanalysierter Rest		1513

Tabelle 3: Extraktionsresultate für das Verb *apprendre*

8 Problemfälle

Probleme, sog. 'false positives', ergeben sich grösstenteils daraus, dass bei einem oberflächensyntaktischen Vorgehen gewisse 'höhere' linguistische Informationen fehlen. So werden z.B. Nomina- und Mehrwortausdrücke falsch extrahiert, da sie nicht als solche, sprich nicht als zusammengehörend erkannt werden. Die eindrücklichsten Beispiele finden sich im Zusammenhang mit der Konstruktion *apprendre*–Infinitiv-Komplement: Was durch die entwickelte Suchroutine als vermeintliches indirektes Objekt erkannt wird, ist in Wirklichkeit das Komplement des Verbs im Infinitiv (Beispiele (6)–(8)).

(6) *L'industrie a appris à faire confiance aux recommandations de cette office.*

(7) *J'espère que le gouvernement apprendra à répondre efficacement à cette demande.*

(8) *Les individus ne peuvent pas apprendre à se conformer aux normes de la justice humaine...*

Da diese Dependenz nicht erkannt wird (und gleichzeitig Belege existieren für die tatsächlich gesuchte Konstruktion *apprendre*–Infinitiv-Komplement–indirektes Objekt, siehe Beispiel (3)), zeigt dieser Problemfall wohl am besten die Schwächen der auf oberflächensyntaktische Muster beschränkten regulären Suchroutinen.

Dasselbe Problem tritt auch bei Komposita (Beispiel (9)) sowie bei adverbialen Präpositionalphrasen (Beispiele (10), (11)) auf, wo die mit der Präposition *à* eingeführten Sequenzen als indirekte Objekte verstanden werden.

(9) *Pourquoi accroître les difficultés pour les jeunes qui devraient apprendre à manier les armes* à feu *avec les précautions nécessaires?*

(10) *Nous apprenons maintenant à évaluer les conséquences* à long terme *de nos actions.*

(11) *C'est un homme que beaucoup d'entre nous ont appris à respecter* au fil des ans.

Leider können die angesprochenen Fehlextraktionen auf dieser linguistischen Abstraktionsebene nicht zufriedenstellend gelöst werden, da die nötigen Subkategorisierungsinformationen für diese Konstruktionen noch nicht vorhanden sind. Auf diese Ausdrücke spezialisierte Lexika müssten in einem der Extraktion vorangestellten Schritt eingesetzt werden, um Mehrwortausdrücke frühzeitig erkennen und dementsprechend markieren zu können, und somit die in diesem Fall erhaltenen Fehlextraktionen zu minimieren.

9 Abschliessende Bermerkungen und Ausblick

Die in diesem Artikel beschriebenen Prozeduren stützen sich auf minimal vorverarbeitete Corpora, welche mit Hilfe regulärer Suchroutinen auf bestimmte Subkategorisierungsmuster Französischer Verben untersucht werden. Durch den Einsatz detaillierter Syntaxlexika kann auf bereits bestehende Listen von Subkategorisierungsmustern zurückgegriffen werden (obwohl diese nicht zwingenderweise vollständig sind!). Dabei gilt es zu beachten, dass die entwickelten Suchroutinen in ihrer Ausdruckskraft klar beschränkt sind, da sie nur auf morphosyntaktische Merkmale zurückgreifen können (siehe Abschnitt 8). Der Vorteil der Arbeit mit regulären Grammatiken liegt jedoch unter anderem darin, dass modular vorgegangen werden kann: einzelne Bausteine der erarbeiteten Suchmuster können beliebig wiederverwendet werden, wie z.B. satzähnliche, durch Kommas begrenzte Einschübe. Daneben enthält jede Suchroutine auch verb- oder verbklassenspezifische Bausteine.

Eine Weiterentwicklung des hier vorgestellten Ansatzes wäre an die Möglichkeit gebunden, 'höhere' linguistische Informationen zu grammatikalischen Funktionen einzubinden und somit genauere (und effizientere) Extraktionsroutinen zu entwickeln. Eine Verfeinerung der Abfragen ist dabei einerseits corpusabhängig und andererseits direkt an den gewünschten Grad an Granularität gekoppelt.

Anmerkungen

[1]Die hier vorgestellten Resultate beziehen sich auf eine Untersuchung der drei Verben *apprendre, connaître* und *savoir* im Rahmen einer Magisterarbeit an der Universität Basel ("La polysémie de trois verbes cognitifs (*apprendre, connaître* et *savoir*) et leur désambiguïsation à l'aide de corpus informatisés"). Die Beispiele beziehen sich jedoch nur auf das Verb *apprendre*.

[2]Kollokationen sind von besonderem Interesse für Lexikographen, da sie die häufigsten Verwendungen eines Lexems illustrieren sowie unabhängige Mehrwortausdrücke aufdecken, die ins Lexikon aufgenommen werden könnten/sollten.

[3]Auf die linguistische Vorverarbeitung kann kein Einfluss mehr genommen werden, was zur Folge hat, dass *Tagging*-Fehler bei der Entwicklung der Suchroutinen sowie bei der Evaluation der erzielten Resultate mitberücksichtigt werden müssen.

[4]Obwohl beide Varianten möglich sind, lässt sich eine tendenzielle Preferenz von indirektem Objekt vor Infinitiv-Komplement ausmachen.

[5]Diese umfassen Infinitiv-Komplemente, *que*-Sätze und indirekte Fragesätze.

Literatur

[Busse and Dubost, 1989] Winfried Busse and Jean-Pierre Dubost. *Französisches Verblexikon. Die Konstruktion der Verben im Französischen.* Klett, Stuttgart, 1989.

[Eckle-Kohler, 1998] Judith Eckle-Kohler. Methods of quality assurance in semi-automatic lexicon acquisition from corpora. In *Proceedings of EURALEX'98*, pages 119–128, Liège, 1998.

[Gahl, 1998] Susanne Gahl. Automatic extraction of subcategorization frames for corpus-based dictionary-building. In *Proceedings of EURALEX'98*, pages 445–452, Liège, 1998.

[Gross, 1986] Maurice Gross. *Grammaire transformationnelle du français*, volume II. Asstril, Paris, 1986.

[Silberztein, 1993] Max Silberztein. *Dictionnaires électroniques et analyses automatique de textes—Le système INTEX*. Masson, Paris, 1993.