

TIGERin – Grafische Eingabe von Benutzeranfragen für ein Baubank-Anfragewerkzeug

Holger Voormann und Wolfgang Lezius
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
<http://www.ims.uni-stuttgart.de/projekte/TIGER>

Zusammenfassung

Dieses Papier beschreibt die Konzeption und Implementation von TIGERin, einem grafischen Front-End für das Baubank-Anfragewerkzeug TIGERSearch. Das TIGERin-System erlaubt eine visuelle Formulierung von Korpusanfragen und ist damit insbesondere für Gelegenheitsanwender sowie zum Erlernen der Anfragesprache von TIGERSearch geeignet.

1 Einführung

Baubanken stellen die Grundlage für zahlreiche computerlinguistische Anwendungen wie die Entwicklung von Parsern oder die Extraktion lexikalischer Information dar (Lezius, 2001). Für das Deutsche stehen die *VerbMobil-Baubank* für gesprochene Sprache (Hinrichs et al., 2000) und das *Negra-Korpus* für geschriebene Sprache (Skut et al., 1998) zur Verfügung. Daneben befindet sich im Projekt *TIGER* eine weitere Baubank im Aufbau, die auf den Ergebnissen des Negra-Projekts basiert und die Annotation von 40.000 Sätzen vorsieht (Brants et al., 2002).

Die Menge an Information, die in einer Baubank annotiert ist, ist ohne speziell entwickelte Werkzeuge nicht verwertbar. Das Suchwerkzeug *TIGERSearch* wurde im Rahmen des TIGER-Projekts entwickelt und verarbeitet Baubanken, deren Annotation in Form von Baumstrukturen oder eingeschränkten Graphstrukturen vorliegt (Lezius, 2002a; Lezius, 2002b). Suchanfragen werden in TIGERSearch in einer formalen Anfragesprache formuliert. Das folgende Beispiel beschreibt Nominalphrasen, die aus Artikel, Adjektiv und Nomen bestehen (vgl. Ergebnisvisualisierung in Abb. 1):

```
#n1:[cat="NP"] &  
#t1:[pos="ART"] & #t1:[pos="ADJA"] &  
#t1:[pos="NN"] &
```

```
#n1 > #t1 & #n1 > #t2 & #n1 > #t3 &  
#t1 . #t2 & #t2 . #t3 &  
arity(#n1,3)
```

Diese Anfrage spezifiziert zunächst den Nominalphrasen-Knoten anhand seiner syntaktischen Kategorie (`cat="NP"`) sowie die drei Token anhand ihrer Wortart Artikel, Adjektiv und Nomen. Anschließend wird ausgedrückt, dass jedes Token direkt von der Nominalphrase dominiert wird (Dominanzsymbol: ">"). Um die Reihenfolge der Token festzulegen, wird zusätzlich die Präzedenz zwischen den Token-Paaren spezifiziert (Präzedenzsymbol: "."). Es bleibt die Festlegung der Stelligkeit der Nominalphrase durch das Prädikat `arity`, da die Phrase ansonsten weitere Token umfassen könnte.

Wie dieses Beispiel zeigt, können Anfragen in einer logischen Anfragesprache schnell komplex und unübersichtlich werden. Für Gelegenheitsanwender, die sich auf die Arbeit mit dem Korpus konzentrieren wollen und weniger am Erlernen der Anfragesprache interessiert sind, ist diese Form des Korpuszugangs nicht geeignet.

2 Grafische Erstellung von Suchanfragen

Als grafisches Front-End für das TIGERSearch-Werkzeug wurde TIGERin entwickelt (Voormann, 2002). Dieses System erlaubt das Zeichnen einer Anfrage. Eine solche Anfrage, die eine unterspezifizierte Graphbeschreibung darstellt, wird aus den Elementarbausteinen Knoten und Relationen aufgebaut. Die grafische Repräsentation orientiert sich an der Ergebnisvisualisierung von TIGERSearch. Durch eine sprechende Symbolik und Eingabemöglichkeit über Menüs bleibt die Darstellung auf das Notwendigste beschränkt. Abbildung 2 zeigt die angegebene Beispielanfrage in grafischer Form.

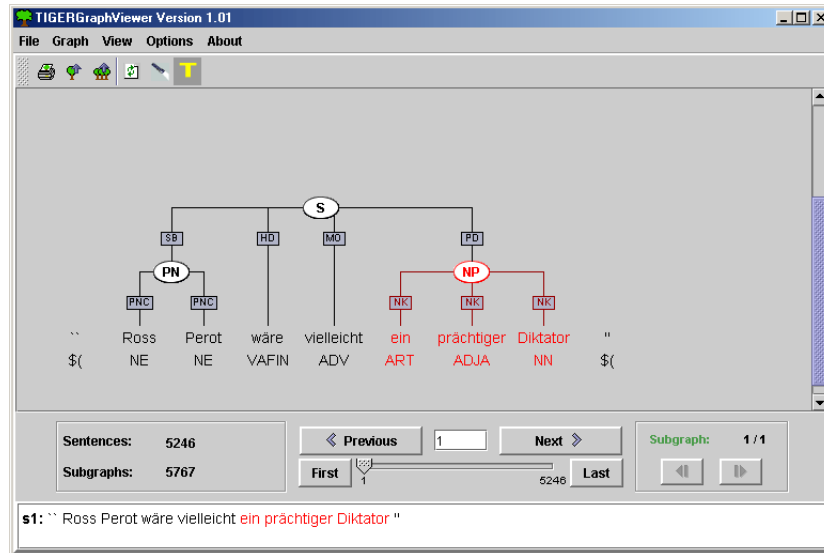


Abbildung 1: Visualisierung eines matchenden Satzes des TIGER-Korpus

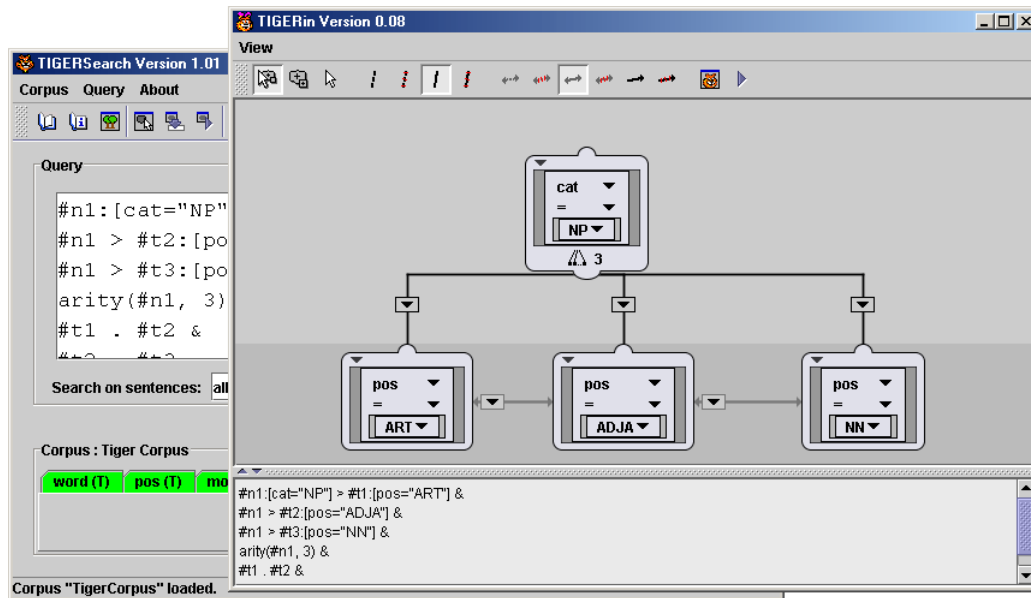


Abbildung 2: Ein Beispiel für eine grafische Suchanfrage

TIGERin präsentiert sich dem Benutzer in einem Fenster, das eine zweigeteilte Arbeitsfläche enthält. Durch einen Mausklick im oberen Bereich wird ein innerer Knoten erzeugt, durch einen Mausklick im unteren Bereich ein Wortknoten (Token). Da Wortknoten üblicherweise auf einer Ebene dargestellt werden, sind sie in TIGERin nur horizontal verschiebbar. Oberhalb der Wortknotenebene lassen sich Knoten frei platzieren und verschieben.

Im inneren, dunklen Bereich eines Knotens wird ein Knoten durch Attribut-Wert-Paare beschrieben. Die Eingabe eines Attribut-Wert-Paars erfolgt in drei Schritten mit drei untereinander angeordneten angeordneten Pull-downmenüs (vgl. Abb. 2). Im ersten Schritt wird das Attribut ausgewählt, im zweiten Schritt der Operator und im dritten Schritt der Attributwert. Als Operatoren dienen u.a. Gleichheit, Ungleichheit und Match gegen einen regulären Ausdruck. Zur Erleichte-

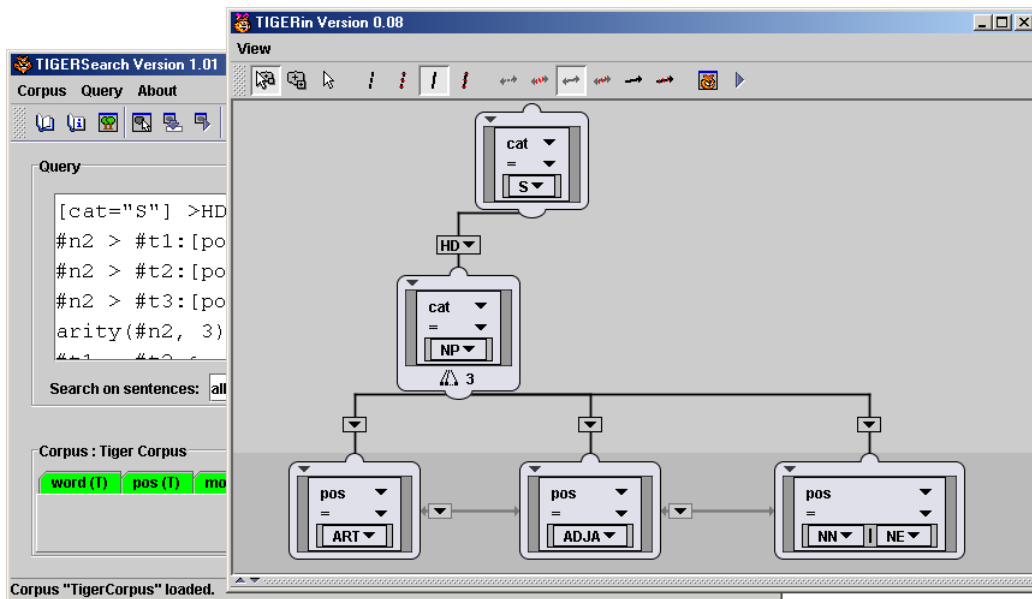


Abbildung 3: Schrittweise Verfeinerung einer Korpusanfrage

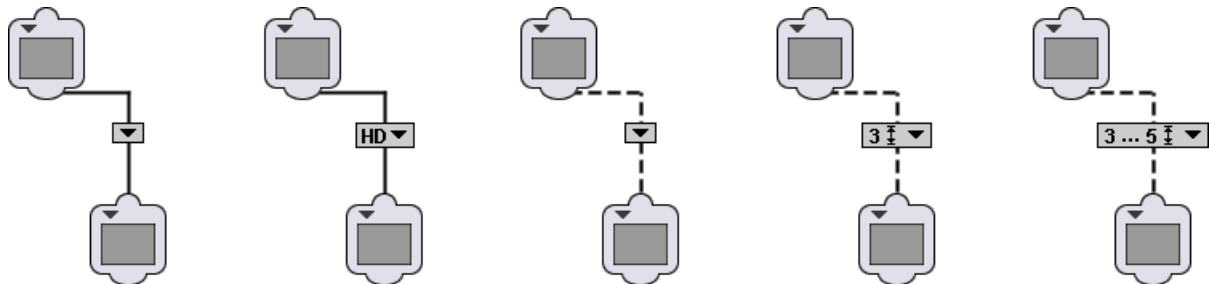


Abbildung 4: Eine Auswahl von Dominanztypen: direkte Dominanz $>$, beschriftete Dominanz $>HD$, allgemeine Dominanz $>*$, n -stufige Dominanz >3 , Dominanz mit Stufenbereich $>3,5$

nung stehen auch Operatoren wie *enthält*, *beginnt mit* und *endet mit* zur Verfügung, die in der generierten Textdarstellung der Anfrage in einen regulären Ausdruck transformiert werden. Ist der Wertebereich eines Attributs bekannt, so kann der Attributwert aus einer Liste möglicher Werte ausgewählt werden.

Zur weiteren Spezifikation eines Graphknotens können Attribut-Wert-Paare miteinander verknüpft werden. Eine zusätzlich eingefügte Attributbeschreibung wird konjunktiv mit der benachbarten Beschreibung verknüpft. Diese Verknüpfung kann per Mausklick in eine Disjunktion verwandelt werden. Analog kann der Wertebereich bei einem Attribut-Wert-Paar angegeben werden. In Abb. 3 kann somit das Wortart-Attribut *pos* des rechten Tokens entweder den Wert "NN" oder "NE" besitzen.

Dominanzrelationen werden durch Erzeugen von Linien zwischen zwei Knoten erstellt. Als Startpunkte dienen die Halbovale an der unteren Kante des Knotens, Endpunkte sind die Halbkreise an der oberen Kante. Der Relationstyp (allgemeine oder direkte Dominanz bzw. jeweils negiert) lässt sich durch das Kanten-Pull-downmenü ändern. Ebenfalls über ein Pull-downmenü werden optionale Kantenbeschriftungen bzw. Entfernungsangaben ausgewählt (vgl. Dominanztypen in Abb. 4). Abbildung 3 zeigt die verfeinerte Variante der Beispielanfrage. Es wird nun ein Nominalphrasentyp beschrieben, der Subjekt eines Satzes ist.

Linien, die sich zwischen den seitlichen Grenzen von Knoten befinden, stellen Präzedenzrelationen dar (vgl. Abb. 2 und 3). Ihre Handhabung verläuft analog zur Dominanzrelation.

Wie alle Kanten besitzen auch Knoten ein Menü, das hier zur Auswahl von Knotenprädikaten wie der Knotenstelligkeit (**arity**) fungiert. Die Menüs ermöglichen den Verzicht auf Dialoge. Einstellungen lassen sich so direkt an den Objekten vornehmen. Gleichzeitig wird nur das angezeigt, was in der textuellen Anfrage eine Entsprechung hat (vgl. Stelligkeitssymbol in Abb. 2 und 3).

3 Implementation

TIGERin ist wie das TIGERSearch-Werkzeug in Java implementiert. Hauptnutzer der Werkzeuge werden Forschungseinrichtungen mit unterschiedlichen und heterogenen Rechnerlandschaften sein. Durch Java ist der Einsatz auf den verschiedensten Plattformen möglich.

Da TIGERin eine objektorientierte Sichtweise verwendet, entsprechen grafische Objekte wie Knoten, Kanten oder Attributsbeschreibungen im Programmcode jeweils Objektklassen. Die Wiederverwendung oder Erweiterung ist durch Vererbung leicht möglich.

4 Verwandte Arbeiten

Bei den meisten verfügbaren Baumbank-Suchwerkzeugen müssen Korpusanfragen auf Kommandozeilen-Ebene gestellt werden. Zu den wenigen Suchwerkzeugen mit grafischer Benutzeroberfläche zählen *VIQTORYA* und *ICECUP*. Im *VIQTORYA*-System kann eine Anfrage mit Hilfe von Dialogen aus Textbausteinen aufgebaut werden (Steiner und Kallmeyer, 2002). Dieser Zugang erleichtert die Formulierung von Suchanfragen, kann aber eine grafische Lösung nicht ersetzen. *ICECUP* bietet sowohl eine grafische Eingabe von Anfragen als auch eine grafische Visualisierung von Anfrageergebnissen an (Wallis und Nelson, 2000). Die in TIGERSearch benötigten Ausdrucksmittel kreuzende und sekundäre Kanten werden jedoch nicht unterstützt.

5 Ausblick

Die Entwicklung von TIGERin ist noch nicht abgeschlossen. Als nächste Schritte sind die Unterstützung von Variablen und die Repräsentation der Disjunktion auf Graphenebene geplant.

Nach der Fertigstellung dieser Arbeiten wird TIGERin ein Bestandteil der Distribution von TIGERSearch sein, um zusätzlich die grafische

Eingabe von Suchanfragen zu ermöglichen. Benutzerstudien sollen dann Aufschluss darüber geben, in wie weit das Eingabewerkzeug von den Benutzern angenommen wird bzw. welche Funktionalität noch fehlt.

Literaturverzeichnis

- Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, und George Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.
- Erhard W. Hinrichs, Juli Bartels, Yasuhiro Kawata, Valia Kordoni, und Heike Telljohann. 2000. The Verbmobil Treebanks. In Werner Zühlke und Ernst G. Schukat-Talamazzini, Herausgeber, *Konvens 2000 Sprachkommunikation*, S. 107–112. VDE-Verlag.
- Wolfgang Lezius. 2001. Baumbanken. In Kai-Uwe Carstensen, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde, und Hagen Langer, Herausgeber, *Computerlinguistik und Sprachtechnologie - Eine Einführung*, S. 377–385. Spektrum Akademischer Verlag, Heidelberg, Berlin.
- Wolfgang Lezius. 2002a. Ein Werkzeug zur Suche auf syntaktisch annotierten Textkorpora. Dissertation, in Vorbereitung.
- Wolfgang Lezius. 2002b. TIGERSearch – Ein Suchwerkzeug für Baumbanken. In *Tagungsband zur Konvens 2002*, Saarbrücken.
- Wojciech Skut, Thorsten Brants, Brigitte Krenn, und Hans Uszkoreit. 1998. A linguistically interpreted corpus of German newspaper texts. In *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*, S. 18–24, Saarbrücken.
- Ilona Steiner und Laura Kallmeyer. 2002. VIQTORYA - A visual query tool for syntactically annotated corpora. In *Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002)*, S. 1704–1711, Las Palmas.
- Holger Voormann. 2002. TIGERin - Grafische Eingabe von Suchanfragen in TIGERSearch. Diplomarbeit, Fakultät Informatik, Universität Stuttgart.
- Sean Wallis und Gerald Nelson. 2000. Exploiting fuzzy tree fragments in the investigation of parsed corpora. *Literary and Linguistic Computing*, 15(3):339–361.