CORTE

Computerlinguistische Methoden für die Rechtsterminologie

DFG PI 154/10-1

Experimente zur Definitionsextraktion aus Gerichtsurteilen

Manfred Pinkal & Stephan Walter

Universität des Saarlandes, Saarbrücken
Computerlinguistik

Projekt CORTE

- Extraktion und Verarbeitung von Definitionen aus Urteilstexten
- Einsatz computerlinguistischer Analyse: Semantisch orientierte Parsing-Technologie (PREDS-Parser, Braun 2003)
- Regelbasierte Vorgehensweise

Überblick

- Motivation: Definitionen
- Beispiel
- Verarbeitung von Rechtssprache
- Experimente und Evaluation
- Fazit und Ausblick

Definitionen in Urteilen

- Rechtsbegriffe im Gesetzestext
  - Legaldefinitionen, Bsp.: *Sache, Geschäft, Vertrag*
  - Unbestimmte Rechtsbegriffe,
    Bsp.: *Treu und Glauben, Stand der Technik*

- 'Zweite Ebene' undefinierter Begriffe in Legaldefinitionen,
  Bsp: *Gewässerbett*

- Normalsprachliche Begriffe im Tatbestand
  - Zunächst meist ohne Definition

- Urteilsbegründung: Vermittlung zwischen diesen Ebenen
  - Verbindung zwischen Gesetz und Sachverhalt
  - Verbindung zwischen Begriffen
  - Verbindung zwischen Urteilen (Zitate und Modifikation)

  => Definitionen haben hier zwar nicht zwangsläufig rechtliche, jedoch hohe praktische Bindungswirkung

Strukturelle Definitonstypen

- Parenthetisch / Appositiv
  - Klammerdefinitionen: *A (sog. B)*
  - Nicht-restriktive Relativsätze:

    Der Begriff der **polizeilichen Gefahr**, *der Wahrscheinlichkeit eines Schadenseintritt*, …
- Parataktisch (meist unmarkiert)

  … **fehlt** der Beschwerde der Antragsgegnerin das **Rechtsschutzbedürfnis**. Sie *kann* ihre *Rechtsstellung im Beschwerdeverfahren nicht mehr verbessern*.
- Prädikat-basierte Definitionen
  - 'Definitor-Verb'+Subjunktion: A *liegt vor, wenn* B
  - Copula: A *ist ein* B*, das* C
  - Nomen + Funktionsverb:
    Die begriffliche Voraussetzung für A *besteht darin, dass* B

=> sprachlich markiert, reiche interne Struktur

Beispiel-Definition

Bei einem Einfamilienreihenhaus liegt *ein mangelhafter Schallschutz dann* vor, wenn *die Haustrennwand einschalig errichtet wurde.*

- ***Definiendum*** */ Definiens*
- Sachbereich
- Verbindung
  / Signalwort

Weitere Möglichkeiten:

- Modifikation der Geltung (z.B. *in der Regel*)
- Zitatangabe
- Rechtsbereich

Computerlinguistische Analyse

Erzeugt normalisierte, Semantik-nahe Struktur

- Verb+Präfix, Argumente
- Modifikation
- Aktiv / Passiv
- Tempus
- Modus / modale Hilfsverben

➔ Abstraktion z.B. über Oberflächenvarianten

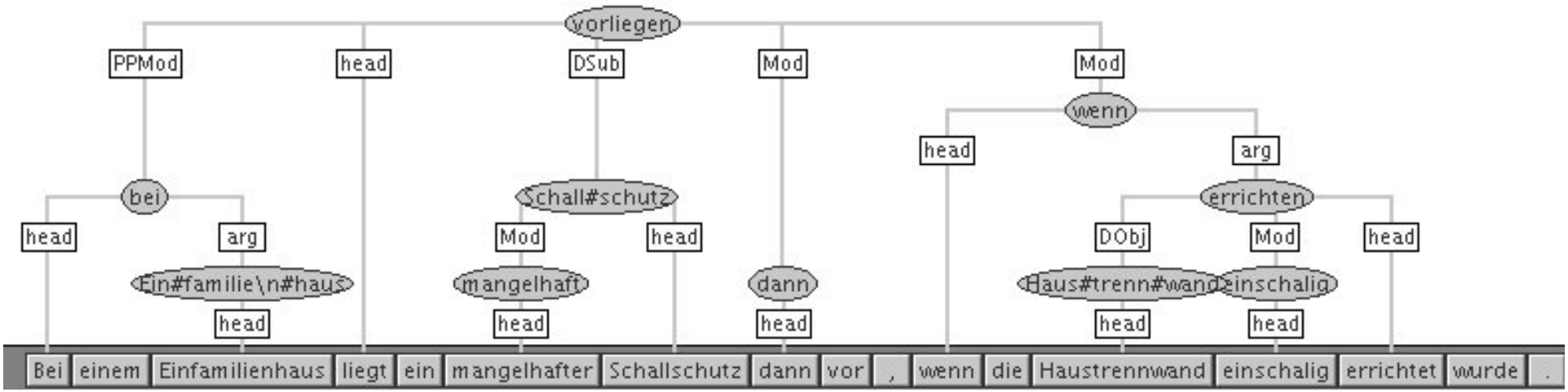
- defined] *liegt vor, wenn* [defining]
- [defined] *vorliegt, wenn* [defining]
- *wenn* [defined] muss [defining] *vorliegen*
- ...

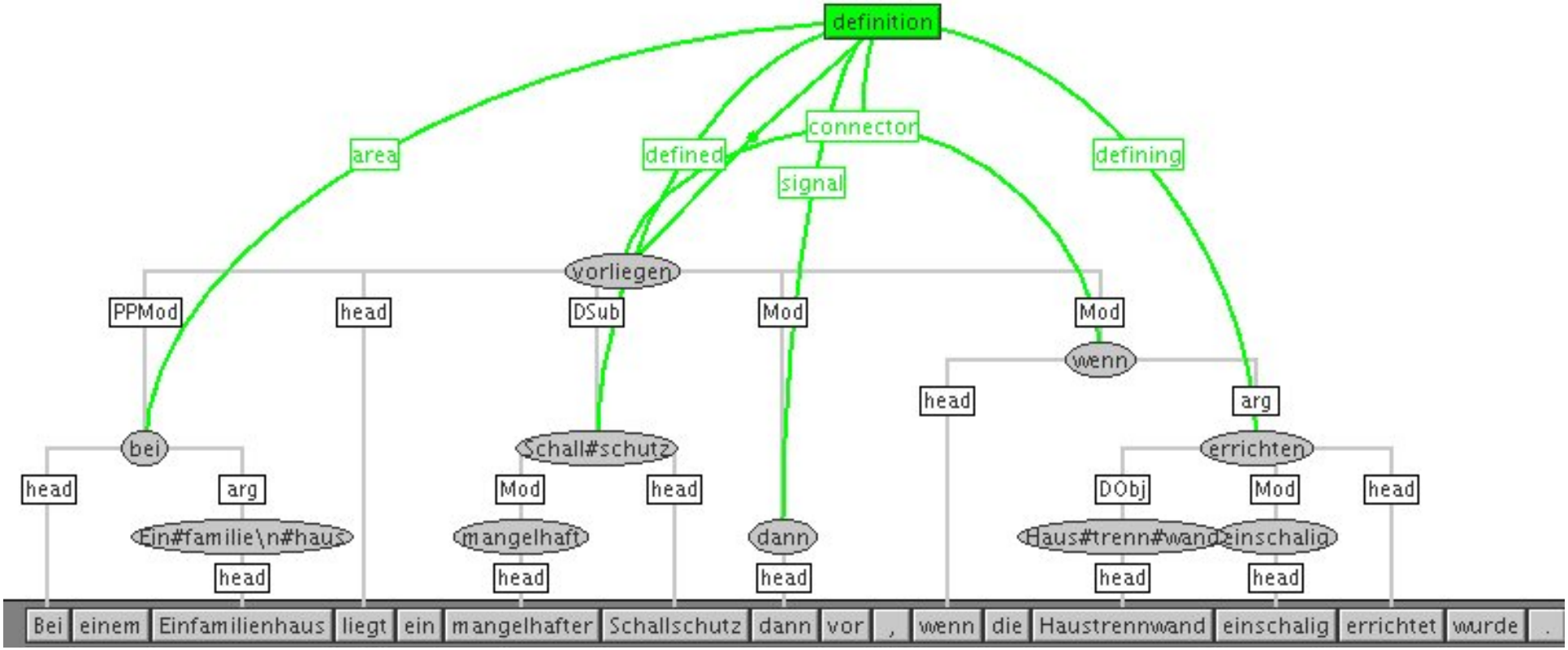
➔ Zugriff auf Definitionsbestandteile

Beispiel Linguistische Struktur

Bei einem Einfamilienreihenhaus liegt ***ein mangelhafter Schallschutz*** dann vor, wenn *die Haustrennwand einschalig errichtet wurde.*

```
vorliegen[verb, sg, ind]
  -PPMod->bei[praep, dat]
            -Arg->*ein#familien#reihen#haus[noun]
  -DSub->*schall#schutz[noun]
           -Mod->mangelhaft[adj]
  -Mod->wenn[subj]
          -Arg->errichten[verb, sg, ind, pres, pass]
                  -DObj->*haus#trenn#wand[noun, defArt]
                  -Mod->einschalig[adv]
  -Mod->dann[adv]
```


vorliegen
PPMod
head
DSub
Mod
Mod
wenn
bei
Schall#schutz
head
arg
errichten
head
arg
Mod
head
DObj
Mod
head
Ein#familie\n#haus
mangelhaft
dann
Haus#trenn#wand
einschalig
head
head
head
head
head
Bei
einem
Einfamilienhaus
liegt
ein
mangelhafter
Schallschutz
dann
vor
,
wenn
die
Haustrennwand
einschalig
errichtet
wurde
.


definition
area
defined
connector
signal
defining
vorliegen
PPMod
head
DSub
Mod
Mod
wenn
bei
Schall#schutz
head
arg
errichten
head
arg
Mod
head
DObj
Mod
head
Ein#familie\n#haus
mangelhaft
dann
Haus#trenn#wand
einschalig
head
head
head
head
head
Bei
einem
Einfamilienhaus
liegt
ein
mangelhafter
Schallschutz
dann
vor
,
wenn
die
Haustrennwand
einschalig
errichtet
wurde
.

Vorgehensweise

- Offline-Vorverarbeitung von Eingabedokumenten:
  - Pre-Processing (Segmentierung, Markup von Nummerierung u.ä., Satzgrenzen)
  - Parsen der Leitsätze, Orientierungssätze, Entscheidungsgründe (Preds-Parser, Braun 2003)
  - ➔ XML-Struktur
- Suche über XPath-Ausdrücke auf Basis des Prädikats (z.B. Top-Node *vorliegen* mit *wenn*-Nebensatz)
- Weiterverarbeitung:
  - Interne Segmentierung der Fundstellen
  - In Vorbereitung: Ablage der Suchergebnisse in einer life abfragbaren Datenbank

Verwendete Textgrundlage

Subkorpus	Urteile	Sätze	Wörter (tokens)	Wörter/Sätze ohne Tatbestand
Umweltrecht 1979 – 2001	9226	568 437	18 164 024	14 342 888 / 418 691
Verwaltungsrecht 1995 – 1997	17 829	934 054	30 382 448	22 383 618 / 638 529
Verschiedene Rechtsbereiche	9884	875 478	26 615 327	21 335 022 / 655 186

Mit Dank an *juris*!

Verarbeitung von Rechtssprache

1. Spezielle Ausdrücke: Abkürzungen, Sonderzeichen, Zitate
2. Sprachliche Komplexität
3. Semantische Besonderheiten

Zitate, Abkürzungen, Spezialausdrücke

- Schwärzungen und Anonymisierung

- Zitatangaben als Satzteile:

  Auf das Asylrecht des Art. 16a Abs. 1 GG kann sich gemäß Art. 16a Abs. 2 Satz 1 GG, § 26a Abs. 1 Sätze 1 und 2 AsylVfG nicht berufen, wer…

=> *Grammatik für Zitatangaben*

- Abkürzungen und Eigennamen:

  Zur Behandlung des Agios durch das FA macht die Astin geltend…

  - Lokal definierte Abkürzungen: Astin => *Antragstellerin*
  - Dazu: Namen und Abkürzungen aus verschiedensten Bereichen (*Firmen, Strassen, Städte, Währungen…*)

=> *Aktuell ca. 45 000 Abkürzungen und Namen erfasst, mehrere tausend davon juristisch*

Komplexität: Präpositionalphrase

[$_0$ *Bei der Umsetzung*
[$_1$ *der Vorgaben*
[$_2$ *der Gerichte*]
[$_2$ *für eine verfassungskonforme Regelung*
[$_3$ *der Überführung*
[$_4$ *von Ansprüchen und Anwartschaften*]
[$_4$ *aus den Zusatz- und Sonderversorgungssystemen*
[$_5$ *der ehemaligen DDR*]
]]
]
]
]

Satzkomplexität - Eckdaten

	Satzlänge	Tiefste Satz-einbettung	Prädikate	Abhängig-keiten
Umweltrecht	*33,32*	*1,65*	*1,93*	*4,26*
Verwaltungs-recht	*33,02*	*1,65*	*1,95*	*4,24*
Zeitung: FR (Tiger-Corpus)	*17,37*	*1,35*	*1,54*	*3,06*

Semantische Komplexität

- Beispiele: Modifikator-Komplexe

*Solche ernstlichen Zweifel liegen **unter anderem immer dann** vor, wenn …*

*… liegt der Versagungsgrund **aber nicht nur dann nicht** vor, wenn …*

Auswertung und Experimente

- Korpus mit hand-annotierten Definitionen
  - 40 Urteile, 4799 Sätze
  - Klassifikation: sicher / unsicher
    - Total: 123 Definitionen
    - 83 ein Satz, 40 zwei bis fünf Sätze lang
    - 87 sicher
    - 90 Prädikat-basiert (93 Sätze)

  => Ausweitung der Datengrundlage in Arbeit
- Umweltrecht-Korpus

Experimente

- Experiment 1
  - 33 Suchmuster
  - Strenge Filterung (z.B. Reihenfolge, Stoppwörter)

=> *Evaluation auf annotiertem Korpus*

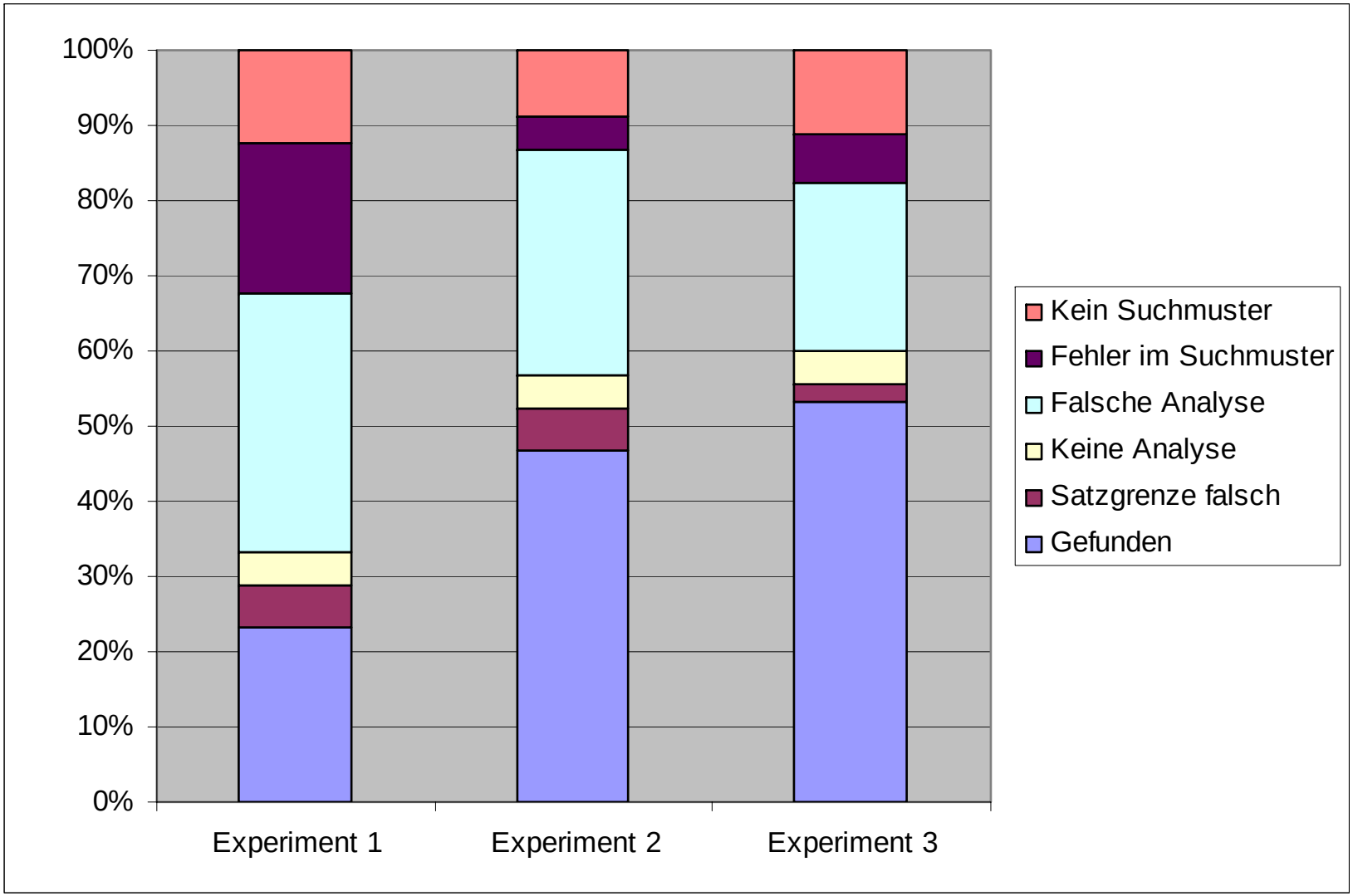
- Experiment 2
  - 59 Suchmuster
  - Keine Filterung

=> *Verbesserung der linguistischen Verarbeitung*

- Experiment 3

Recall

- Wie viele der Ziel-items (d.h. der tatsächlich vorhandenen Definitionen) wurden gefunden
- Auswertung auf annotiertem Korpus

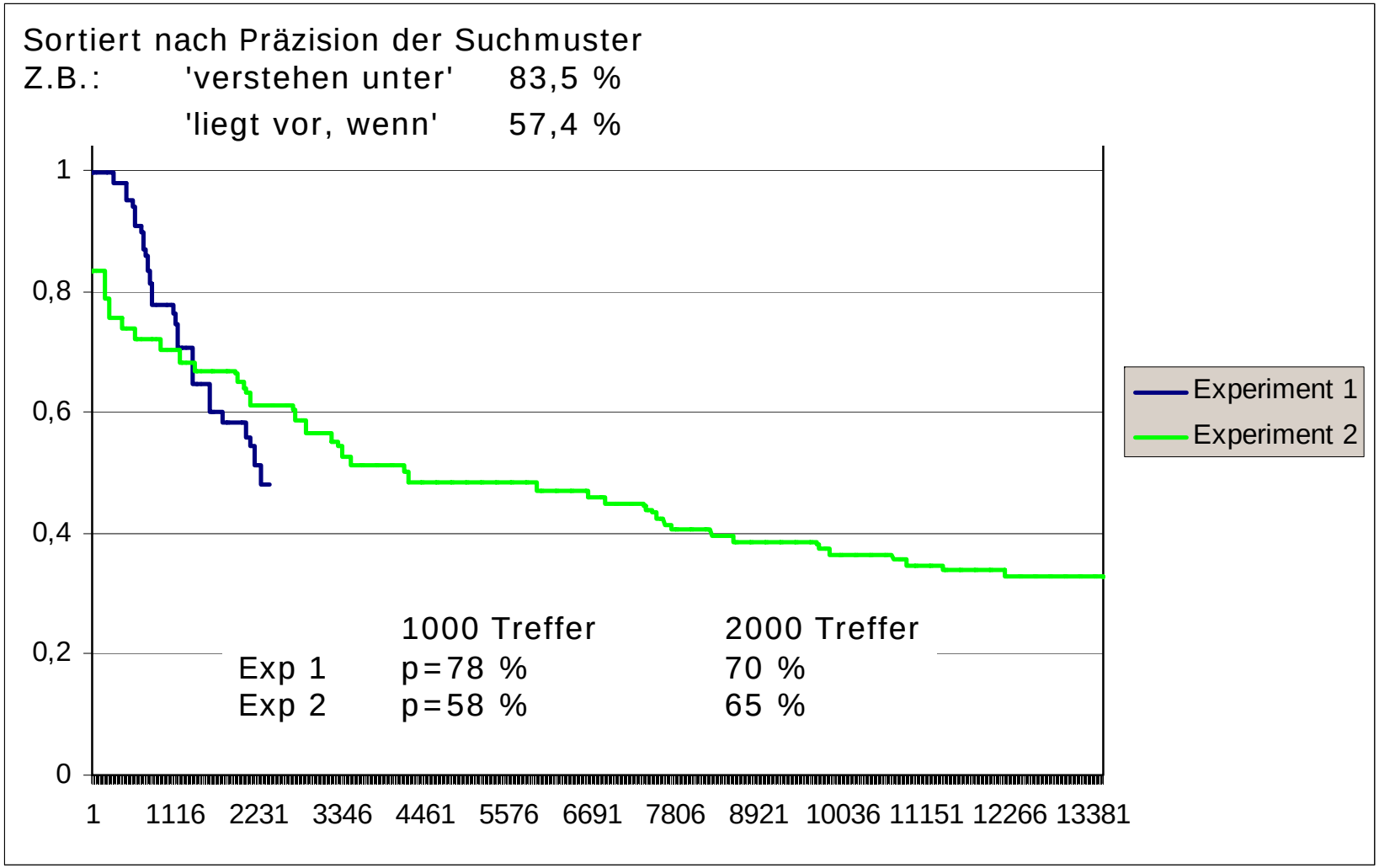


Gefunden

Experiment 1	21
Experiment 3	48

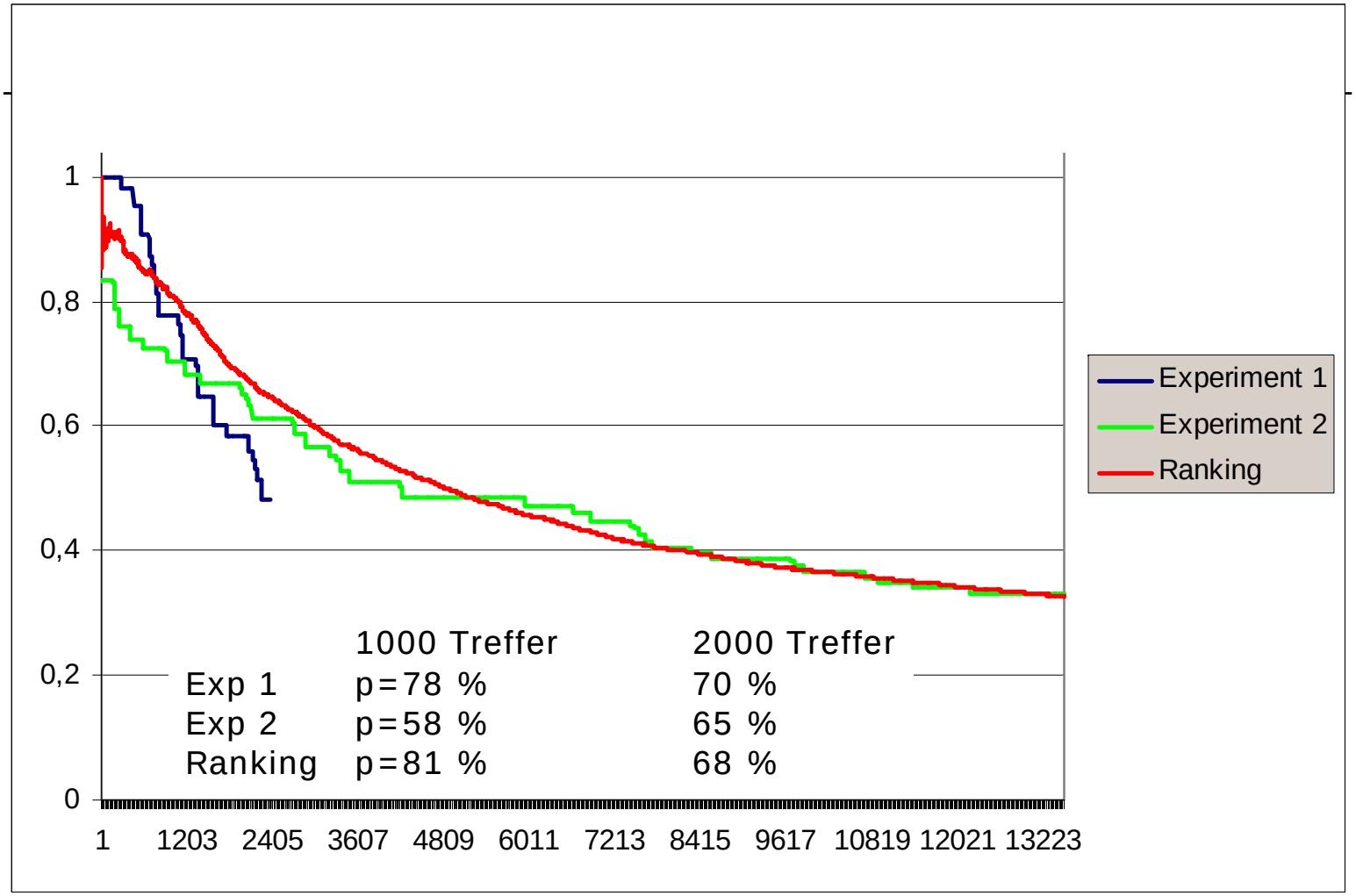
Präzision

- Wie groß ist der Anteil an *true positives* an den Suchergebnissen*?*
- Auswertung auf Umweltrechtkorpus
- Bewertung von ca. 1500 Treffern (Experiment 1) und ca. 3800 Treffern (Experiment 2)
- Experiment 3 noch nicht durchgeführt (Parser läuft noch…)


Sortiert nach Präzision der Suchmuster
Z.B.: 'verstehen unter' 83,5 %
'liegt vor, wenn' 57,4 %
1
0,8
0,6
0,4
0,2
0
1 1116 2231 3346 4461 5576 6691 7806 8921 10036 11151 12266 13381
1000 Treffer
2000 Treffer
Exp 1 p=78 % 70 %
Exp 2 p=58 % 65 %
Experiment 1
Experiment 2

Ranking

- Flexible Kombination *weicher* Kriterien statt *alles-oder-nichts*-Filterung
- Zum Beispiel:
  - Stopp- und Signalwörter
  - Definitheit / Verallgemeinerungs-Signale
  - Wortstellung (Definiendum vor Definiens)
  - Lexikalische Ähnlichkeit mit bekannten Definitionen
- Gewichtung automatisch gelernt (Regressionsanalyse)


1
0,8
0,6
0,4
0,2
0
1
1203
2405
3607
4809
6011
7213
8415
9617
10819
12021
13223
Experiment 1
Experiment 2
Ranking
1000 Treffer
2000 Treffer
Exp 1 p=78 % 70 %
Exp 2 p=58 % 65 %
Ranking p=81 % 68 %

Fazit

- Präzise Extraktion interessanter Trefferzahlen möglich
- Automatisch erworbene Parameter ermöglichen Balance von Recall und Präzision
- Problematische Abhängigkeit von der Qualität der linguistischen Verarbeitung

Ausblick

- Evaluation auf größerer Datengrundlage
- Integration von Suchergebnissen
  - Auflösung von Unvollständigkeit
  - Relation zwischen Definienda
  - Präzisierung, Widerspruch, Differenzierung...
- Datenbank / Benutzerinterface
- Fall-Back System mit geringerem Anspruch an linguistische Verarbeitung