The increased availability of electronic text data requires new technologies for extracting relevant information

INFORMATION EXTRACTION (IE)

The goal of IE research is to build systems that find and link relevant information from NL text while ignoring extraneous and irrelevant information

The core functionality of an IE system is quite simple:

Input:

1. Specification of the relevant information in form of templates (feature structures), e.g., company information, product information, management succession, meetings of important peoples
2. A set of real-world text documents

Output:

A set of instantiated templates filled with relevant text fragments (eventually normalized to some canonical form)

Source: GN

Example Information Extraction 1

Lübeck (dpa) - Die Lübecker Possehl-Gruppe, ein im Produktions-, Handel- und Dienstleistungsbereich tätiger Mischkonzern, hat 1994 den Umsatz kräftig um 17 Prozent auf rund 2,8 Milliarden DM gesteigert. In das neue Geschäftsjahr sei man ebenfalls „mit Schwung“ gestartet. Im 1. Halbjahr 1995 hätten sich die Umsätze *des Konzerns* im Vergleich zur Vorjahresperiode um fast 23 Prozent auf rund 1,3 Milliarden erhöht.

type	=	turnover
c-name	=	Possehl1
year	=	1994
amount	=	2.8e+9DM
tendency	=	+
diff	=	+17%

type	=	turnover
c-name	=	Possehl1
year	=	1995/1
amount	=	1.3e+9DM
tendency	=	+
diff	=	+23%

Source: GN

Example Information Extraction 2

Parts from RWE‘s Anual Report (1998):

Eine Schwerpunktregion im Rahmen der Internationalisierung im Energiebereich ist Osteuropa. Hier haben wir unser Engagement im abgelaufenen Geschäftsjahr weiter ausbauen können.Nach dem Kauf weiterer Anteile halten *wir* inzwischen jeweils knapp über 50% an den ungarischen Energieversorgungsunternehmen ELMÜ, ÉMÁSZ und MÁTRA. Im Falle von MÁTRA hat RWE Energie im April 1998 Anteile an Rheinbraun abgegeben. Die Präsenz in Polen wurde durch Kooperationsvereinbarungen mit den Regionalversorgern Zaklad Energetyczny Krakow S.A. (ZEK) und Stoleczny Zaklad Energetyczny S.A. (STOEN) ....im Frühjahr 1998 weiter ausgebaut.

Group/Subs.	YEAR	KIND	FROM	TO	POT	AMOUNT
RWE	1998	+	ELMÜ			>50%
RWE	1998	+	ÉMÁSZ			>50%
RWE	1998	+	MÁTRA			>50%
RWE Energie	1998	-	MÁTRA	Rheinbraun	4.1998	

Source: GN

From the viewpoint of natural language processing (NLP), IE is attractive for many reasons, including

- Extraction tasks are well defined
- IE uses real-world texts
- IE poses difficult and interesting NLP problems
- IE needs systematic interface specification between NL and domain knowledge
- IE performance can be compared to human performance on the same task

⇒ IE systems are a key factor in encouraging NLP researchers to move from small-scale systems and artificial data to large-scale systems operating on human language (Cowie & Lehnert, 1996)

Source: GN

IE has a high application impact

- IE and information retrieval: construction of sensitive indices which are more closely linked to the actual meaning of a particular text
- IE and text classification: getting fine-grained decision rules
- IE and text mining: improve quality of extracted structured information
- IE and data-base systems: improve semi-structured DB approaches
- IE and knowledge-base systems: combine extracted information with KB (e.g., ontology extraction)
- IE and question/answering systems: fine-grained question grammars

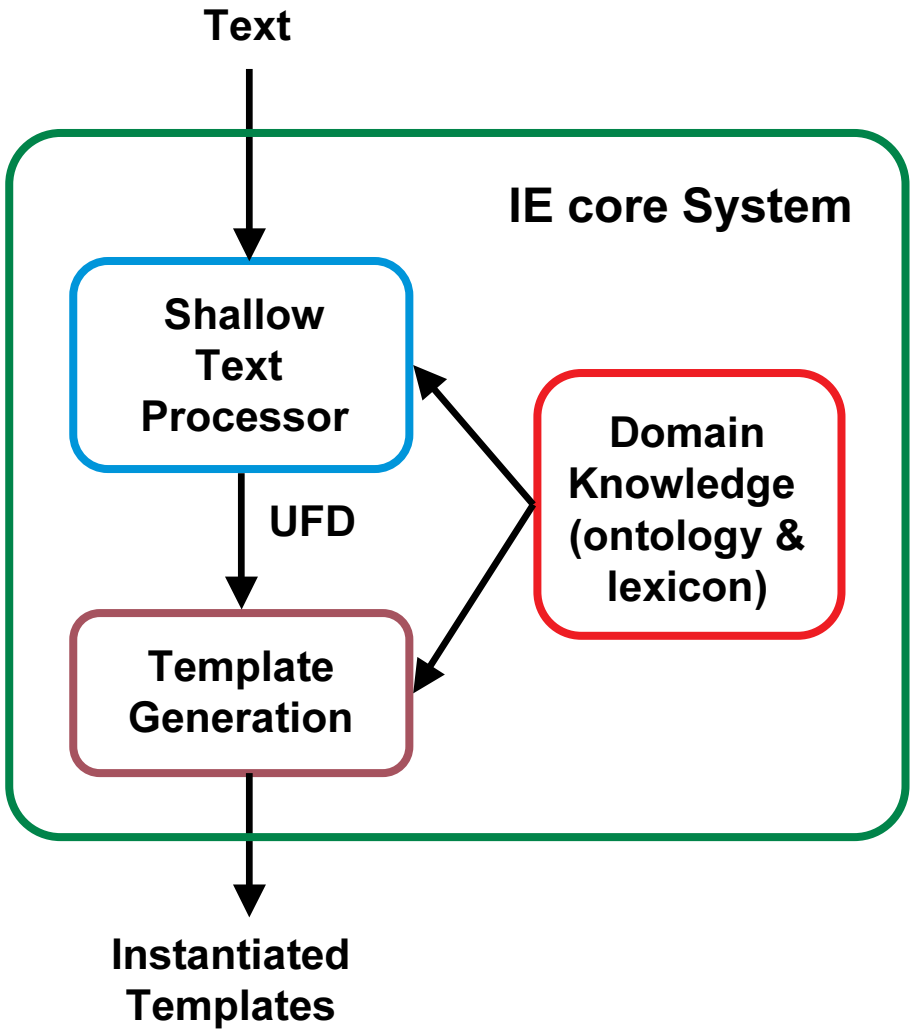
Source: GN

At DFKI we aim at the development of IE-core technologies which can fastly be configured for new application domains and tasks (i.e., fast application development cycle)

MAIN GOALS

- systematic treatment of domain-independent and domain-specific knowledge
  - robust and fast shallow text processor (mildly deep ☺)
  - abstract level of linking between linguistic and domain knowledge

- (semi)-automatic knowledge adaptation to specialized tasks
  - Lexicon & Subgrammar extraction
  - Machine Learning of template filler rules



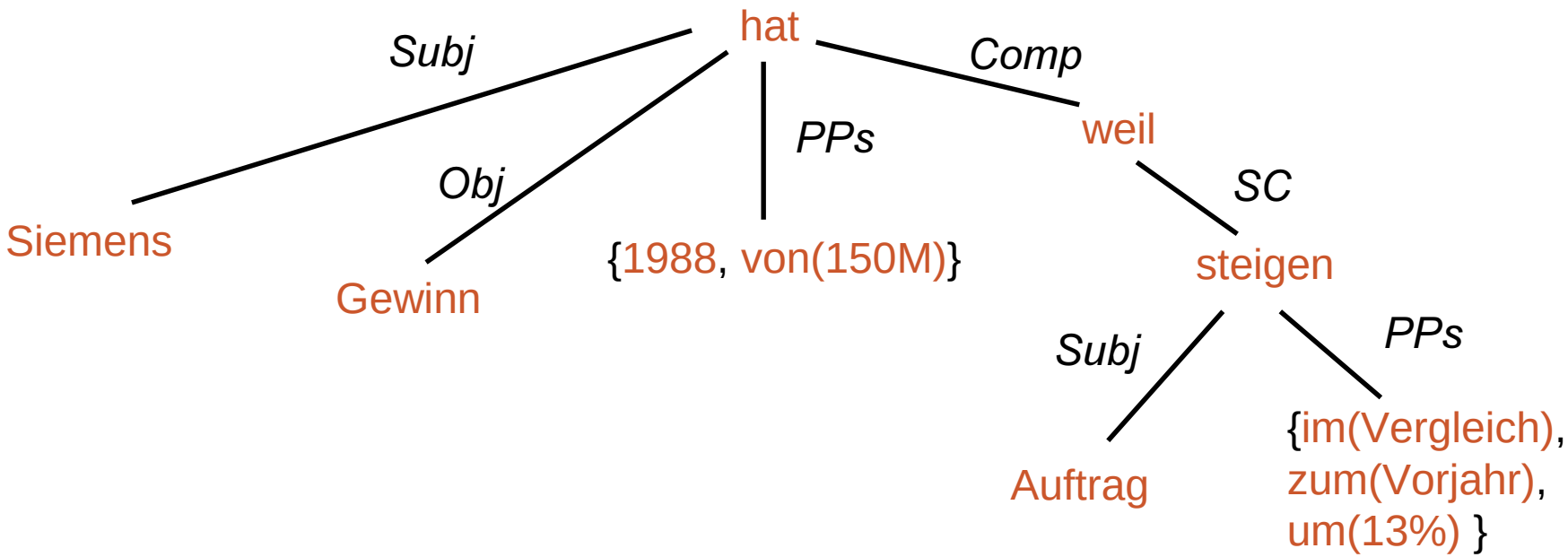
Source: GN

An analysed text is represented as a sequence of underspecified (partial) functional descriptions UFDs

UFD: flat dependency-based structure, only upper bounds for attachment and scoping

[$_{PN}$Die Siemens GmbH] [$_{V}$hat] [$_{year}$1988][$_{NP}$einen Gewinn] [$_{PP}$von 150 Millionen DM],
[$_{Comp}$weil] [$_{NP}$die Auftraege] [$_{PP}$im Vergleich] [$_{PP}$zum Vorjahr] [$_{Card}$um 13%] [$_{V}$gestiegen sind].

"The siemens company has made a revenue of 150 million marks in 1988, since the orders increased by 13% compared to last year."



Source: GN

The major components of the shallow text processor are realized on top of two basic tools

- C++ Toolkit for weighted finite state transducers (WFST)
  - e.g., determination, minimization, concatenation, local extension (following advanced approaches developed at AT&T, Xerox)
  - compact and efficient internal representations

- Dynamic tries for lexical & morphological processing
  - recursive traversal (e.g., for compound & derivation analysis)
  - robust retrieval (e.g., shortest/longest suffix/prefix)

- Parameterizable XML-output interface

- Both tools are portable across different platforms (Unix & Linux & Windows NT)

Source: GN

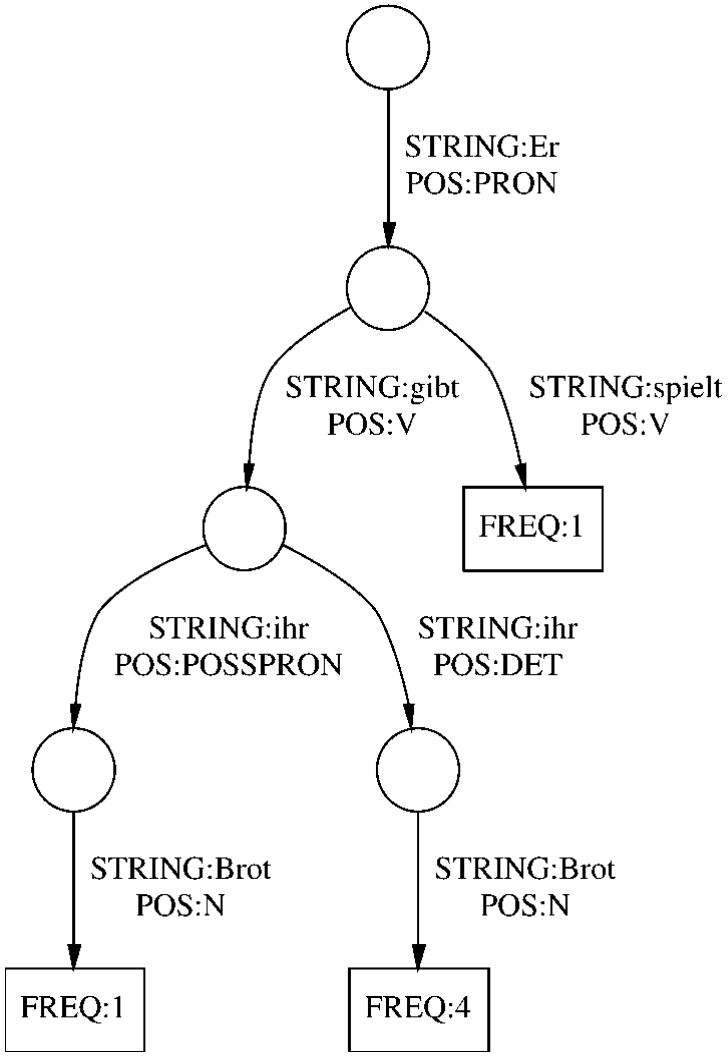
We have developed a C++ toolkit for weighted FSM (cf. Piskorski)

- architecture and functionality is mainly based on the tools developed by AT&T (letter transducer)
- most of the provided operations are based on recent approaches (Mohri, Pereira, Roche, Schabes)
- some of the operations restricted to subclass of FSMs due to limited closure properties of finite-state transducers: determinization, intersection, removing epsilon transitions
- new algorithms (relevant for shallow processing), modifications
  - local extension (adapted for the case of WFST)
  - direct incremental construction of minimal deterministic acyclic finite-state automata
  - improved algorithm for epsilon removal

Source: GN

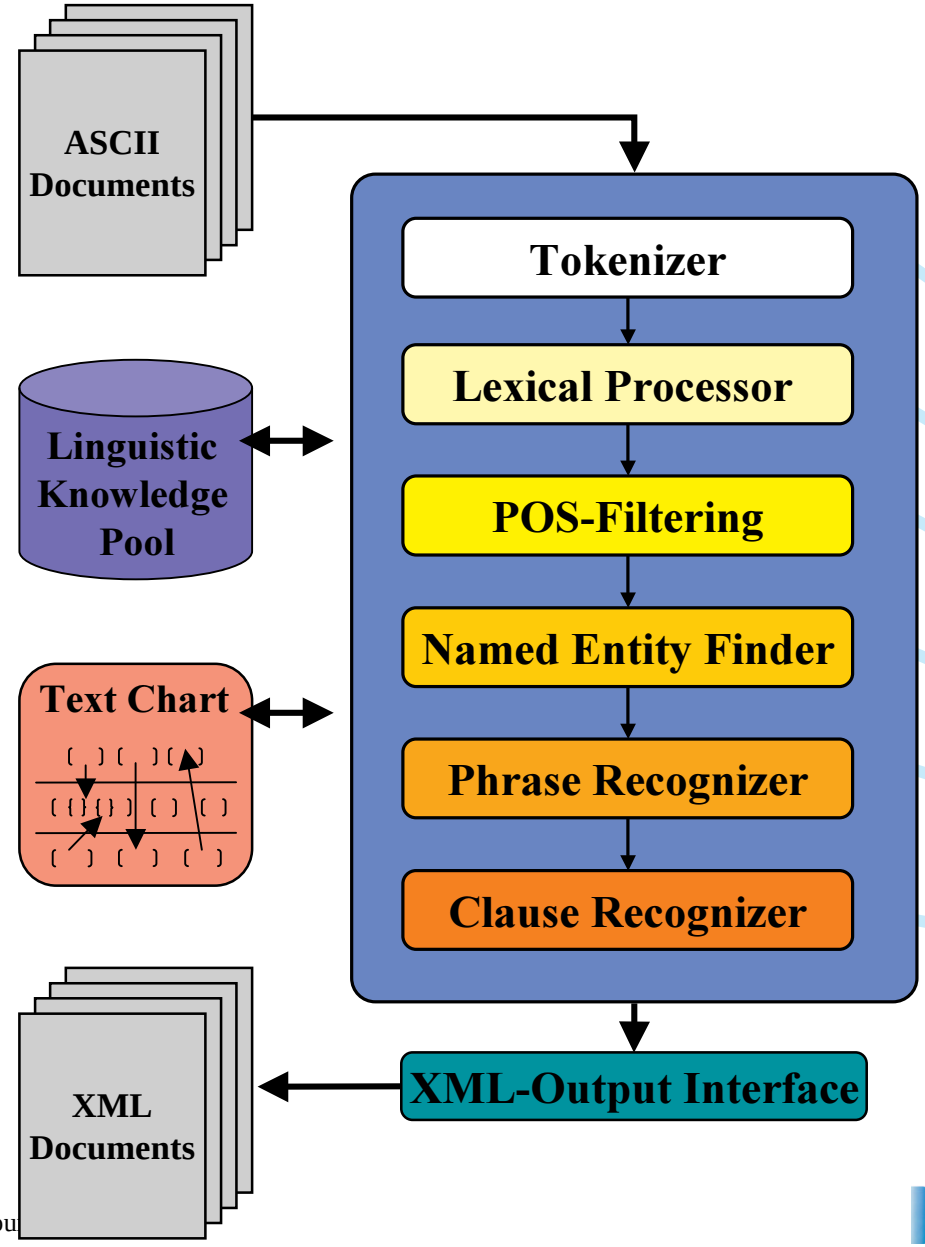
Generic Dynamic Tries

- parameterized tree-based data structure for efficiently storing sequences of elements of any type, where each sequence is associated with an object of some other type (GDT)
- efficient **deletion** function is provided (self-organizing lexica)
- variety of complex functions relevant to linguistic processing supporting recognition of **longest and shortest prefix/suffix** of a given sequence in the trie
- example: Trie for storing verb phrases and their frequencies, where each component of the phrase is represented as a pair <POS,STRING>



Source: GN

At DFKI´s LT-lab we have developed powerful domain-independent shallow text processing components in order to support a fast IE system development cycle



EXAMPLE: rund 60 bis 70 Prozent der Steigerungsrate
(about 60 to 70 percent increase)

rund: low-w
60: 2int

52 classes

Steigerungsrate: steigerung+[s]+rate
bis: prep|adv

750.000 word forms
on-line compounds
hyphen coordination

bis: adv

Over 100 Rules (also Brill's TBL),
Schabes & Roche approach

rund 60 bis 70 Prozent: percentage-NP

12 subgrammars
dynamic lexicon
reference resolution

rund 60 bis 70 Prozent: NP
der Steigerungsrate: NP

Generic NPs, PPs, VGs

Divide-and-Conquer chunk parser

parametrizable

Finite-state technology

Sou

Named Entity Finder

- The task of the NAMED ENTITY FINDER is the identification of:
  - entities: organizations, persons, locations
  - temporal expressions: time, date
  - quantities: monetary values, percentages, numbers
- Identification in two steps:
  - recognition patterns expressed as WFSA are used to identify phrases containing potential candidates for named entities
  - additional constraints (depending on the type of a candidate) are used for validating the candidates and an appropriate extraction rule is applied in order to recover the named entity

  example: „von knapp neun Milliarden auf über 43 Milliarden Spanische Pesetas“
  from almost nine billions to more than 43 billions spanish pesetas

  TYPE: monetary
  SUBTYPE: monetary-prepositional-phrase

- Longest match strategy

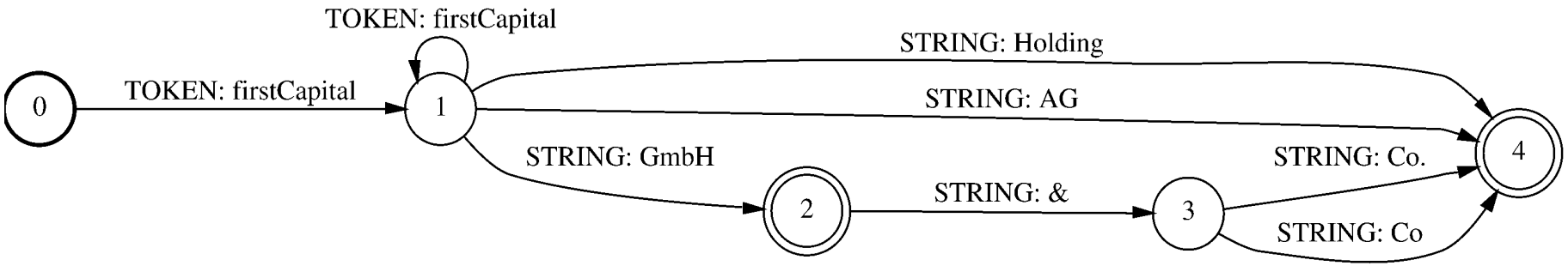
Source: GN

Named Entity Finder (cont.)

- Arcs of the WFSAs are predicates on lexical items:

  (a) **STRING: s**, holds if the surface string mapped by current lexical item is of the form **s**
  (b) **STEM: s**, holds if: the current lexical item has a prefered reading with stem **s** or the current lexical item does not have prefered reading, but at least one reading with stem **s**
  (c) **TOKEN: x**, holds if the token type of the surface string mapped by current lexical item is **x**

- Example: simple automaton for recognition of company names



additional constraint: disallow determiner reading for the first word
candidate: „Die Braun GmbH & Co.“ extracted: „Braun GmbH & Co.“

Source: GN

Named Entity Finder (cont.)

- Additional lexica for geographical names, first names (persons) and company names compiled as WFSA (new token classes)
- Named entities may appear without designators (companies, persons)
- Dynamic lexicon for storing named entities without designators
- Candidates for named entities, example:

Da flüchten sich die einen ins Ausland, wie etwa der Münchner Strickwarenhersteller ***März GmbH*** *oder der badische Strumpffabrikant Arlington Socks, GmbH. Ab kommendem Jahr strickt* ***März*** *knapp drei Viertel seiner Produktion in Ungarn*.

- Resolution of type ambiguity using the dynamic lexicon:
  if an expression can be a person name or company name (*Martin Marietta Corp*.)
  then use type of last entry inserted into dynamic lexicon for making decision

Source: GN

We achieved very good performance and coverage for shallow text processing of German

- **Basis**

  corpus of German business magazine „Wirtschaftswoche“ (1,2MB, 197118 tokens)

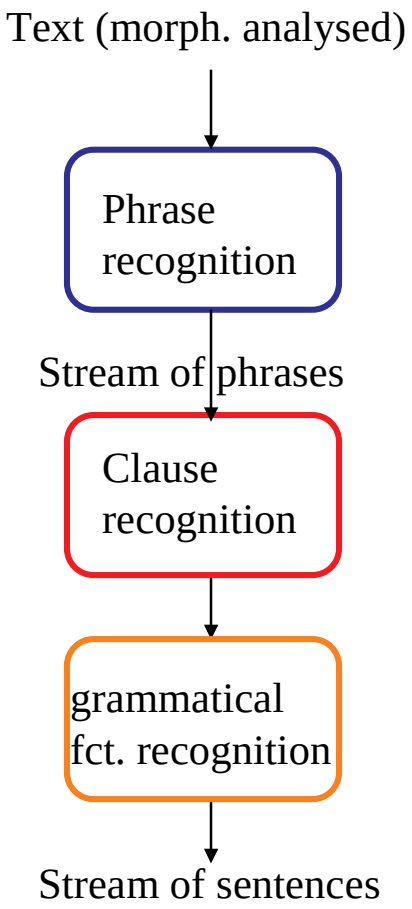
- **Performance**

  ~32sec. (~6160 wrds/sec; PentiumIII, 500MHz, 128Ram)

	Recall	Precision
Evaluation (20.000 tokens)		
– compound analysis:	98.53%	99.29%
– part-of-speech-filterung:	74.50%	96.36%
– Named entity (including NE reference resolution; all 85% R, 95.77% P)		
• person names:	81.27%	95.92%
• companies:	67.34%	96.69%
• locations:	75.11%	88.20%
• total:	73.94%	94.10%
– fragments (NPs, PPs):	76.11%	91.94%

Source: GN

We have identified the needs for better chunk parsing strategies in order to improve robustness and coverage on the sentence level



Current chunk parser

bottom-up:
first phrases and then sentence structure

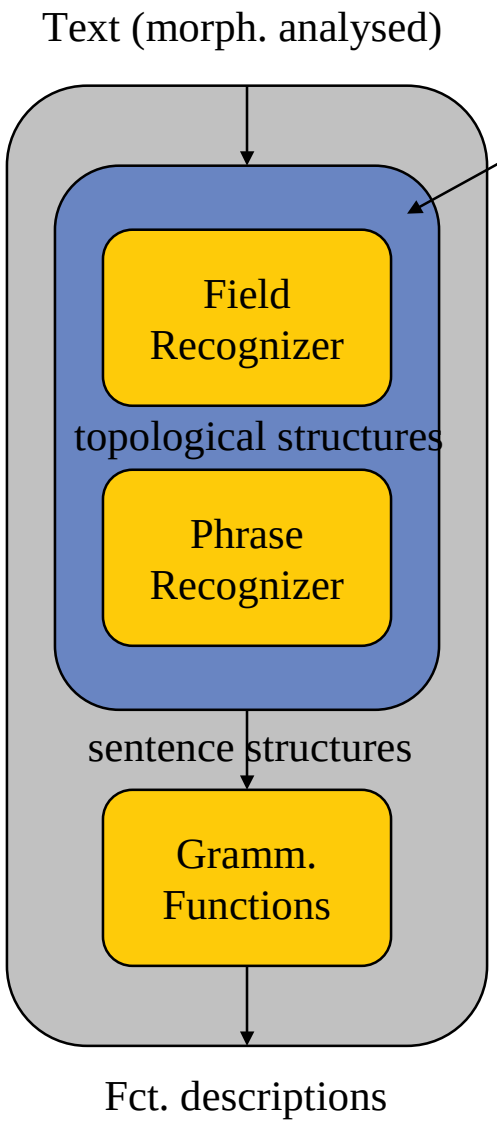
main problem:
even recognition of simple sentence structure depends on performance of phrase recognition

example:
- complex NP (*nominalization style*)
- relative pronouns

[Die vom Bundesgerichtshof und den Wettbewerbern als Verstoss gegen das Kartellverbot gegeisselte zentrale TV-Vermarktung] ist gängige Praxis.
([central television marketing censured by the German Federal High Court and the guards against unfair competition as an act of contempt against the cartel ban] is common practice)

Source: GN

A new chunk parser has been developed that increases robustness and coverage on the sentence level



Divide-and-conquer strategy

first compute topological structure of sentence
second apply phrase recognition to the fields

[coord [core *Diese Angaben konnte der Bundesgrenzschutz aber nicht bestätigen*], [core *Kinkel sprach von Horrorzahlen,* [relcl *denen er keinen Glauben schenke*]].
(This information couldn‘t be verified by the Border Police, Kinkel spoke of horrible figures that he didn‘t believe.)

Evaluation

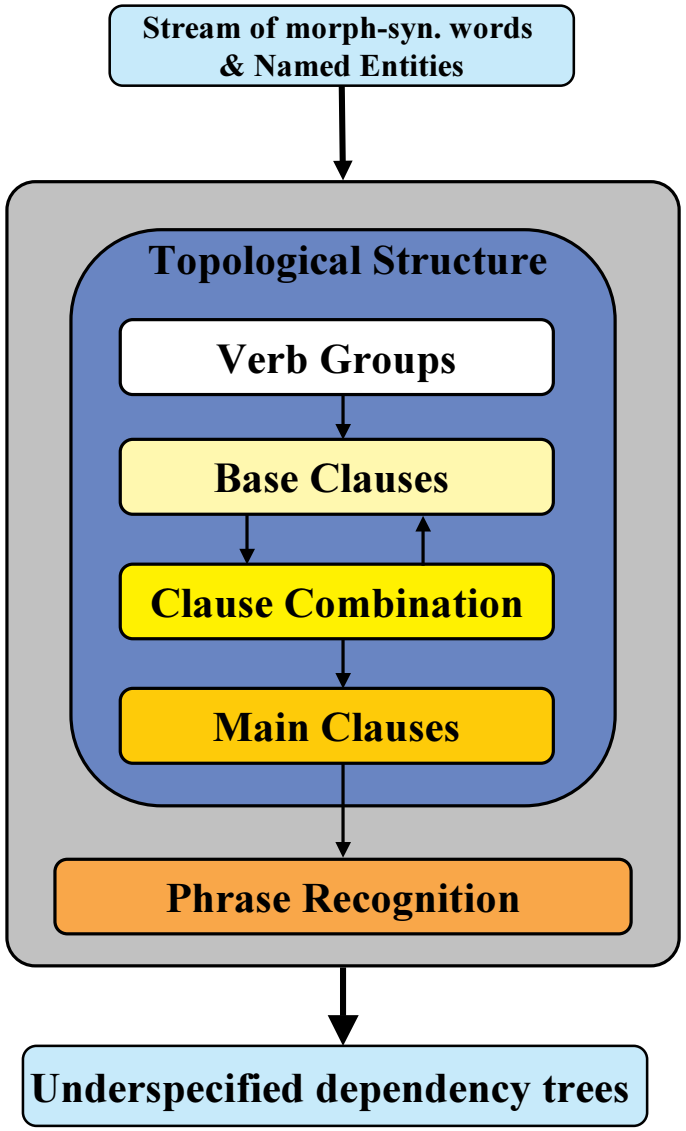
400 sentences (6306 words)

Verb groups:	98.59% F
Clause struct.:	91.62% F
All components:	87.14% F

(more details in Master thesis of C. Braun and NeumannBraunPiskorski:ANLP00)

Source: GN

The divide-and-conquer parser is realized by means of a series of finite state grammars



Weil die Siemens GmbH, die vom Export lebt, Verluste erlitt, mußte sie Aktien verkaufen.
Because the Siemens Corp which strongly depends on exports suffered from losses they had to sell some shares.

Weil die Siemens GmbH, die vom Export Verb-FIN, Verluste Verb-FIN, Modv-FIN sie Aktien FV-Inf.

Weil die Siemens GmbH, Rel-Clause Verluste Verb-FIN, Modv-FIN sie Aktien FV-Inf.

Subconj-Clause, Modv-FIN sie Aktien FV-Inf.

Clause

Source: GN

In order to deal with embedded clauses, two sorts of recursions are identified

Middle-field recursion

embedded base clause is located in the middle field of the embedding sentence

…, weil die Firma, nachdem sie expandiert hatte, größere Kosten hatte.

(*…, *because the company, after it expanded had, increased costs had*.)

- …, weil die Firma [Subclause], größere Kosten hatte.
- … [Subclause].

Rest-field recursion

embedded clause follows the right verb part of the embedding sentence

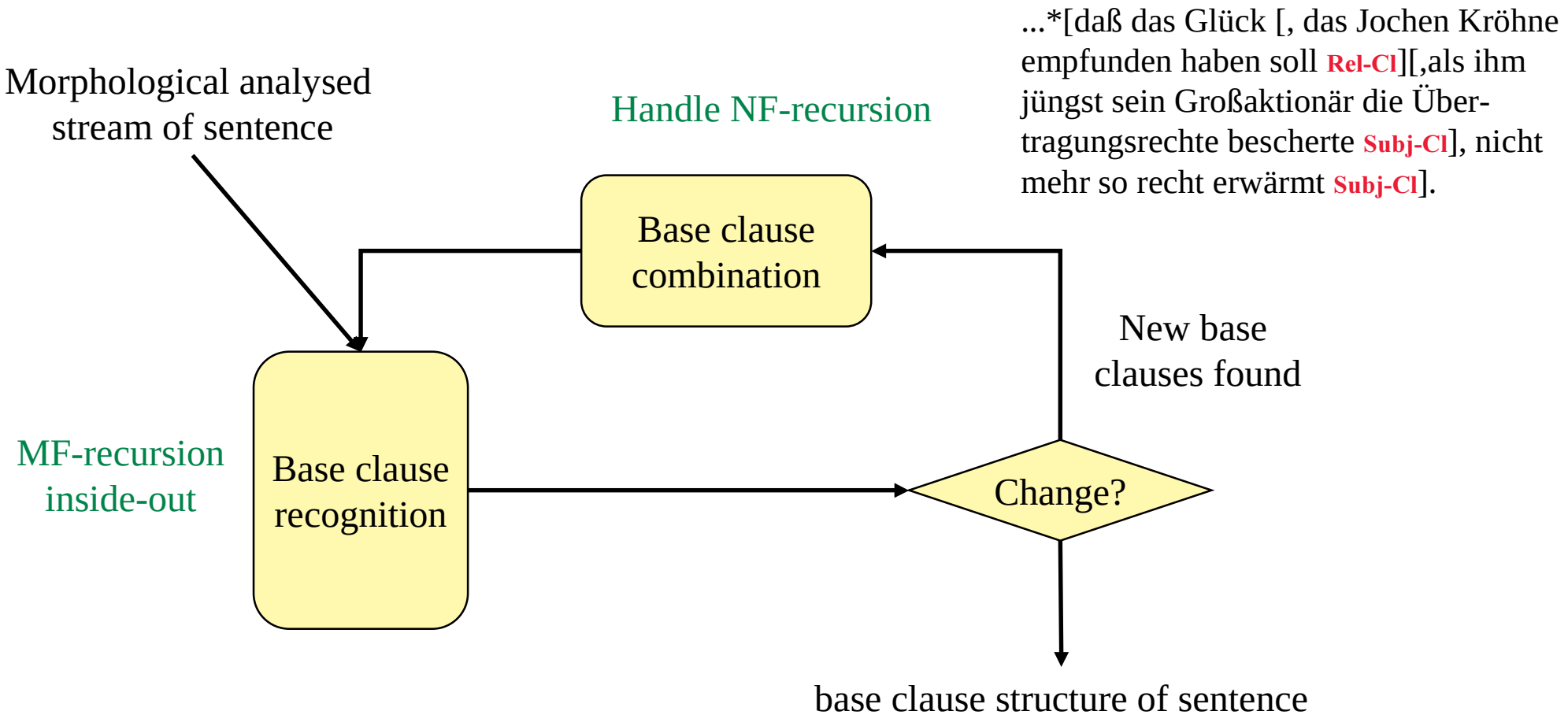
…, weil die Firma größere Kosten hatte, nachdem sie expandiert hatte.

(*…, *because the company increased costs had, after it expanded had*.)

- … [Subclause] [Subclause].
- … [Subclause].

Source: GN

These recursions are treated as iterations which destructively substitute recognized embedded base clauses with their type



...*[daß das Glück [, das Jochen Kröhne empfunden haben soll **Rel-Cl**][,als ihm jüngst sein Großaktionär die Über-tragungsrechte bescherte **Subj-Cl**], nicht mehr so recht erwärmt **Subj-Cl**].

Source: GN

Evaluation on unseen test data (press releases)

Divide-and-conquer parser (400 sentences, 6306 words)

verb module	98.10	98.43	
base-clause module	93.08 (94.61)	93.80 (93.89)	
main-clause module	89.00 (93.00)	94.42 (95.62)	
complete analysis	84.75	89.68	F=87.14

Source: GN

The divide-and-conquer approach offers several advantages

Improved robustness

topological sentence structure determined on basis of simple indicators like verbgroups and conjunctions and their interplay;

phrases need not be recognized completely

Resolution of some ambiguities

relative pronouns vs. determiners

subjunction vs. preposition

clause vs. NP coordination

Modularity

easy exchange/extension of (domain-specific) phrase grammars

Some more examples (source text)

topological structure

plus expanded phrase structure

Source: GN

What we learnt concerning shallow parsing

Divide-and-conquer parsing strategy

free German text processing

suited for free worder languages

high modularity

Main experience

full text processing necessary even if only some parts of a text are of interest;

application-oriented depth of text understanding;

the difference between shallow and deep NLP seen as a continuum

Source: GN

Shallow nominal reference resolution is needed in order to improve template merging

- Goal: find different verbalizations of the same entity

- Example:

  Da flüchten *sich die einen* ins Ausland, wie etwa *der Münchner Strickwarenhersteller* ***März GmbH*** oder *der badische Strumpffabrikant Arlington Socks, GmbH*. Ab kommendem Jahr strickt ***März*** knapp drei Viertel seiner Produktion in Ungarn.

  (*Therefore some take refuge abroad, like the Münchner knitware producer* ***März GmbH*** *or the badische Strumpffabrikant Arlington Socks, GmbH.. From next year on,* ***März*** *knits around three quarters of its production in Hungary.)*

- The design of the reference resolution module is
  - directed by a corpus analysis
  - taking into account various theoretical insights of distinct binding theories (GB and HPSG) or discourse theories (DRT)

Source: GN

Shallow nominal reference resolution module (cf. T. Declerck)

- Modular approach: handle nominal reference problems with actual available structural information as early as possible on different processing levels

- Allows to integrate reference resolution into our cascaded text processing approach
  - compound analysis

    Rohstoff- und Handelsunternehmen => Rohstoffunternehmen und Handelsunternehmen

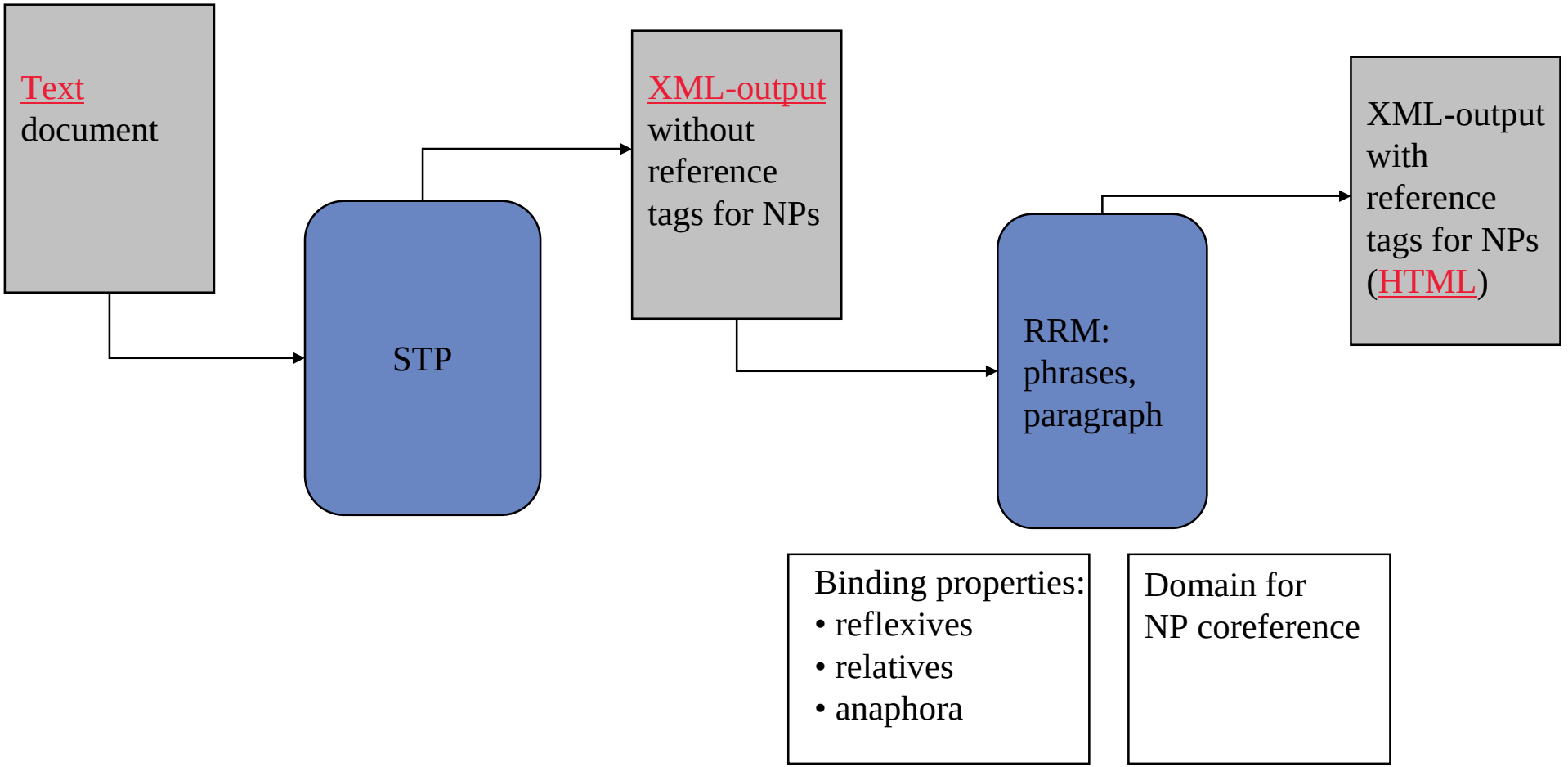
    (extractive and commercial enterprise)

  - named entities
    - [... Mannesmann Mobilfunk GmbH] ([MMO])
    - [...Martin Marietta Corp. ...] ... Marietta

  - phrasal fragments
    - [Abhängige Gesellschaften], [die] das Grössenkriterium nicht erfüllen... *(Dependent societies, which do not fulfill the requirements)*
    - plädierte ... [Roland Issen] für eine flexible Lösung. [Er] will ...; *(pleaded .. Roland Issen for a flexible solution. He wants ...)*

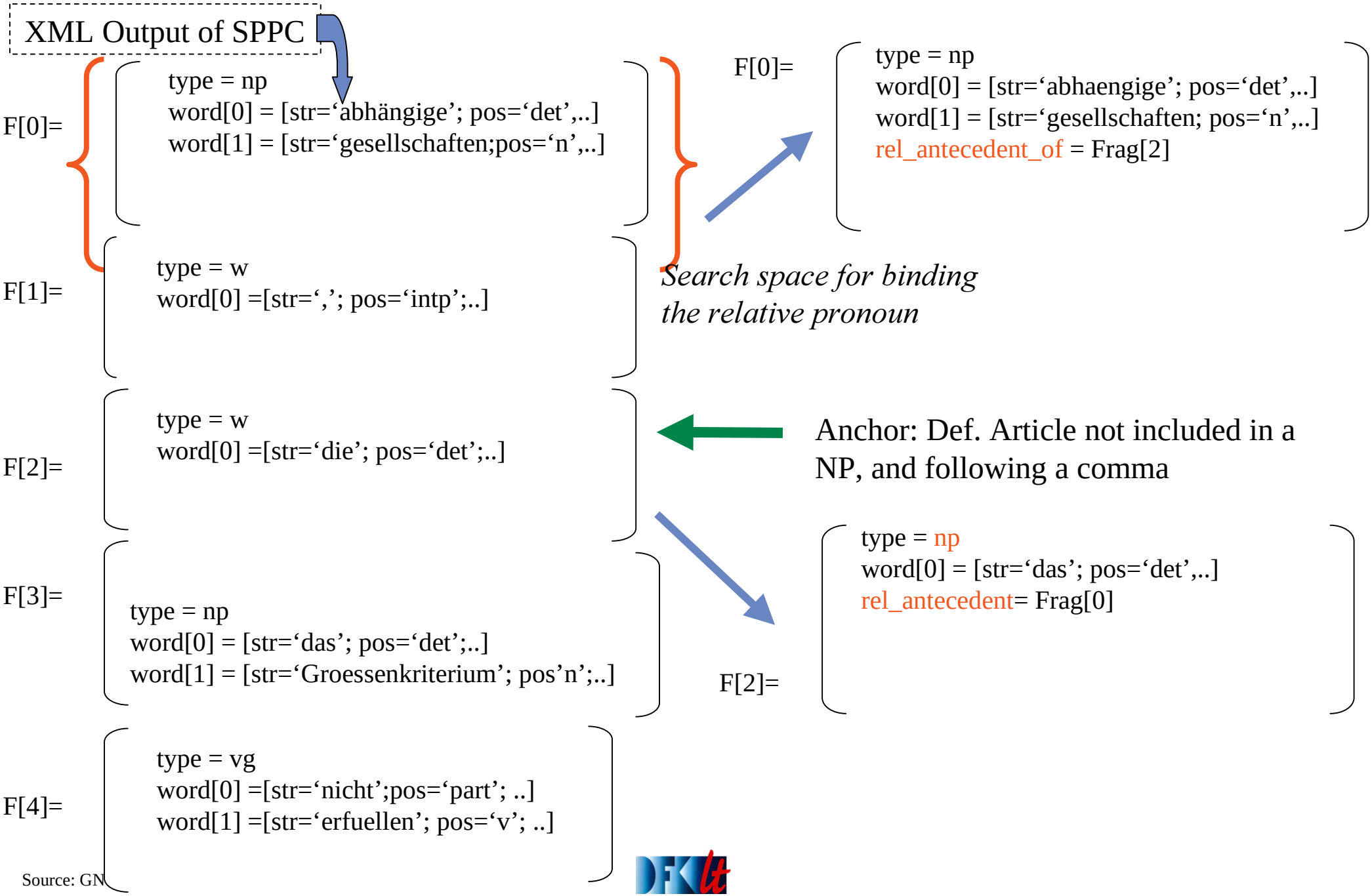
Source: GN

The reference resolution module (RRM) operates on the XML-output of shallow text processor



Source: GN

An example of Pronoun resolution: relative pronoun



Source: GN

Start conclusion

- Robust & efficient software platforms for shallow free text processing of German
  - efficient tools for weighted finite state transducers (comparable to that of Xerox and AT&T)
  - powerful lexical components (incl. compound analysis, POS-filtering)
  - sophisticated named entity recognition (pattern-based, dynamic lexicon)
  - novel chunk parsing strategies (particularily suited for free word order languages)
  - parameterizable XML-interface

- Detailed evaluation on all processing levels

- Marked-up linguistic resources

- Shallow nominal reference resolution

- Multilinguality: Common English & German version under way (joined work together with GETESS, cf. M. Becker)

Source: GN

Conclusion (2)

- Strategies for systematic integration of domain-knowledge
  - defining domain-specific templates as typed feature structure (using TDL formalism, orginally developed for HPSG-development)

- Building of several IE-applications on basis of the proposed model
  - turnover
  - violation of peace treaty
  - soccer world championship 1998

- This topic has already been looped up in other projects
  - GETESS (*German Text Exploitation and Search System*, BMBF-funded, joined project with Uni. Karlsruhe, , Uni. Rostock, Gecko GmbH)
  - Number of DFKI EC projects (Mumis, Muchmore, Airforce, Memphis)
  - Even dot coms: XtraMind, SemanticEdge

Source: GN

The Conclusion

- Paradime finished with concrete results and systems for German free text processing

- Promising starting points for more advanced real-world NLP research
  - Integration of deep processing (BMBF project Whiteboard)
  - Machine Learning for IE (running diploma thesis by V. Morbach)
  - Open domain question answering systems (preparation for qa-trec, foreseen diploma thesis)

Source: GN