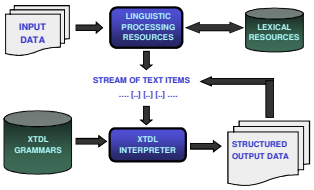
SProUT – Shallow Text Analysis from NER to RE

Stephan Busemann
DFKI GmbH

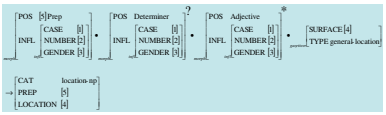
SProUT – Multilingual Information Extraction From Free Text

ARCHITECTURE

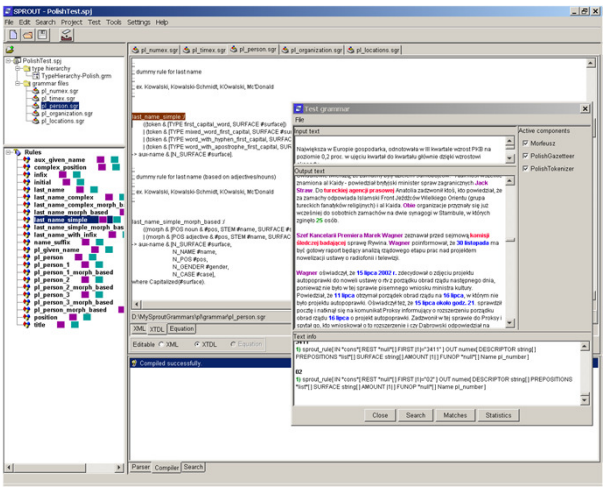


XTDL GRAMMAR FORMALISM

regular expressions over typed feature structures with coreferences and functional application



INTEGRATED DEVELOPMENT ENVIRONMENT



SProUT Processing Components

TOKENIZATION

- Identifying elementary units
- Organized in appx. 30 types, some of them language-specific
- E.g., analysis of currency expressions, such as € 375,000

```
currency_sign : €
comma : ,
any_natural_number: 375
```

MORPHOLOGY

- Relating inflected word forms to word stems and morpho-syntactic features

*MMORPH analysis for the German word form **rasten***

```
rasten -> [POS verb, STEM rasen, NUM pl, PER 1_3, TENSE past]
       -> [POS verb, STEM rasten, NUM pl, PER 1_3, TENSE pres]
       -> [POS verb, STEM rasten, VFORM infinitive]
       -> [POS noun, STEM raste, NUM pl, GEN fem, CASE ngda]
       -> [POS noun, STEM rast, NUM pl, GEN fem, CASE ngda]
```

- Large coverage for 9 languages, including Chinese, Japanese, Polish

GAZETTEER

- For storing static named entities or keywords
- Allows for associating entries with a list of arbitrary attribute-value pairs
- Expressions of the same meaning share a non-linguistic CONCEPT feature across languages, allowing for a homogeneous output

```
cote d'ivoire | GTYPE:gaz_country | CONCEPT:"ivory_coast"
côte d'ivoire | GTYPE:gaz_country | CONCEPT:"ivory_coast"
elfenbeinkueste | GTYPE:gaz_country | CONCEPT:"ivory_coast"
elfenbeinkuste | GTYPE:gaz_country | CONCEPT:"ivory_coast"
elfenbeinküste | GTYPE:gaz_country | CONCEPT:"ivory_coast"
ivory coast | GTYPE:gaz_country | CONCEPT:"ivory_coast"
```

- Reusage of resources for Germanic languages (first names, locations, dates, organizations)
- Specific gazetteers covering terminology
- Acquisition of language specific resources from the Web, and semi-automatic production of all orthographic and morphological variants

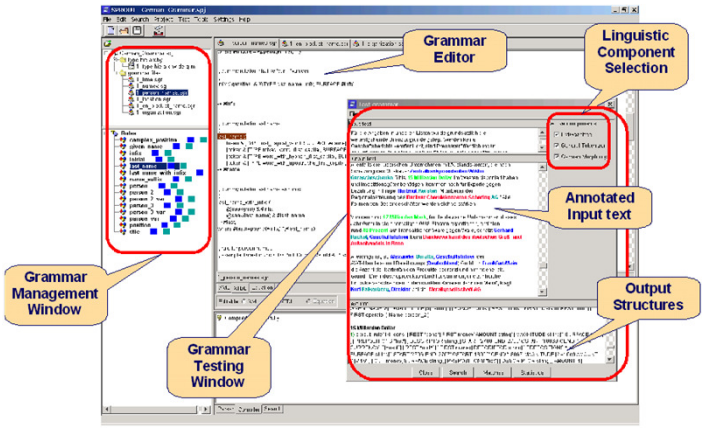
Separate Maintenance, Combined Usage

Rule Component

- Pattern-Action Rules expressed in XTDL combine results of tokenizer, morphology and gazetteer to extract complex elements
  - *Patterns* match the input text sequence
  - *Actions* generate template-like output
- Sample rule matching a *Person Name* followed by a *Function*
  - Person Name
    - Title (optional)
    - Firstname
    - Lastname
    - Comma Function (optional)
  - Function
    - Role
    - Zero to two tokens
    - Country or Organization

	Ex. 1	Ex. 2	Ex. 3
Title (optional)		Dr.	Prof.
Firstname	Barack	Peter	Meinhard
Lastname	Obama	Müller	Briegel
Comma Function (optional)	,	,	
Role	President	Ministerpräsident	Vorstandssprecher
Zero to two tokens	of the	des	der
Country or Organization	USA	Saarlandes	Coralix AG

Integrated Grammar Development Environment



SProUT Supports Basic NER and (some) RE

- **Classical Named Entity Recognition**
  - **Person names, Organizations, Locations, Dates, Currencies, …**
    „Peter Karmanos“, „DFKI GmbH“, „Salt Lake City“, „2009/07/08“, „€ 123.456,78“
- **Extended Named Entity Recognition**
  - **Person names with titles and functions, …**
    „Chairman and CEO of Compuware Dr. Peter Karmanos“
    „Barack Obama, President of the United States“
- **Relation Extraction**
  - **Combination of entities, e.g. Person and its functions over time, yielding a career overview („Dossier“). Can analyze real-world texts like**
    „Vines left his position as Vice President - Communication Services for Chrysler LLC in early December after returning to the company in December 2003. Prior to that, Vines served as a Vice President - External Affairs for Nissan North America from April 1998 until February 2000. From 1993 until 1995 while an employee of Chrysler Corporation, he served as Public Relations Executive with the American Automobile Manufacturers Association. From 1995 to April 1998 he held a variety of labor relations, domestic/international public relations, speechwriting and internal communication positions with Chrysler Corporation.”

Some Technicalities

- Implemented in Java
- Java API, XML RPC
- Lingware for multiple languages at different levels of sophistication
  - de, en, ja, zh, fr, es, nl
- Multiple projects at DFKI, continous maintenance