Exercise #8

Introduction to Knowledge Management, WS2016

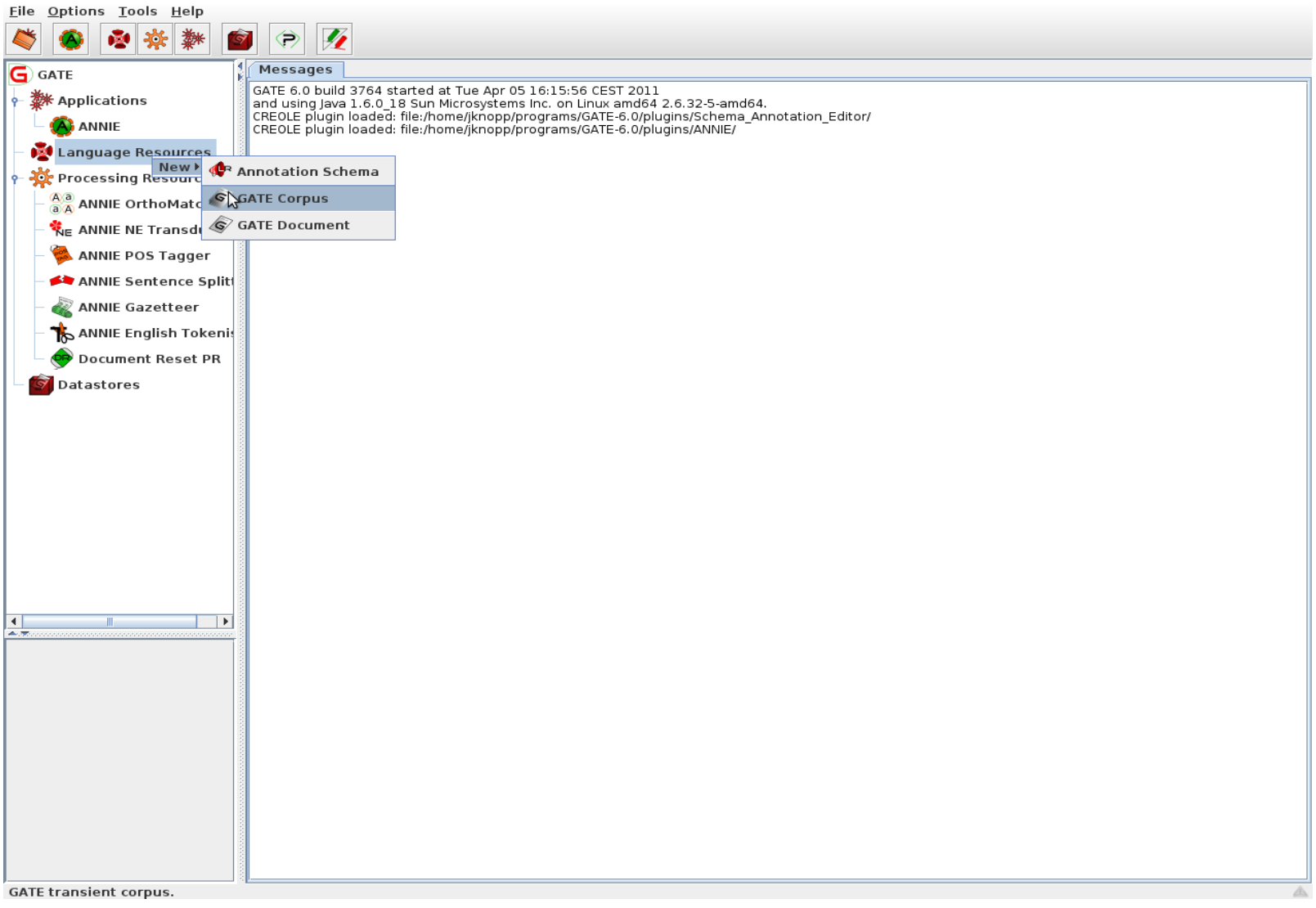
Camilo Thorne

(based on slides by T. Szyler and J. Knopp)

Mannheim, 02.05.2016

Using GATE

1. Load Corpus
2. Run ANNIE
3. Analyze Annotations


File Options Tools Help
GATE
Applications
ANNIE
Language Resources
New
Processing Resources
ANNIE OrthoMatcher
ANNIE NE Transducer
ANNIE POS Tagger
ANNIE Sentence Splitter
ANNIE Gazetteer
ANNIE English Tokeniser
Document Reset PR
Datastores
Annotation Schema
GATE Corpus
GATE Document
Messages
GATE 6.0 build 3764 started at Tue Apr 05 16:15:56 CEST 2011
and using Java 1.6.0_18 Sun Microsystems Inc. on Linux amd64 2.6.32-5-amd64.
CREOLE plugin loaded: file:/home/jknopp/programs/GATE-6.0/plugins/Schema_Annotation_Editor/
CREOLE plugin loaded: file:/home/jknopp/programs/GATE-6.0/plugins/ANNIE/
GATE transient corpus.

Incorrect Annotations

Brown

- Ban Ki-Moon – Person
- Mirek Topalenek – Person
- Shadow – not an Organization
- G20 – Organization

Japan

- $24.6bn – Money
- 3.526tn yen – Money
- 20 trillion yen – Money
- finance ministry – Organization

Lookup Types

Inspect the words annotated with *Lookup* by clicking on the *Annotation List* tab.

List 10 values that are assigned to the *majorType* attribute!

Lookup Types

Inspect the words annotated with *Lookup* by clicking on the *Annotation List* tab.

List 10 values that are assigned to the *majorType* attribute!

facility_key, loc_key, address, number, time_modifier, date_unit, title, time, country_adj person_first, organization, location, currency_unit, jobtitle, stop, cdg,

Lookup Types

Inspect the words annotated with *Lookup* by clicking on the *Annotation List* tab.

List 10 values that are assigned to the *majorType* attribute!

facility_key, loc_key, address, number, time_modifier, date_unit, title, time, country_adj person_first, organization, location, currency_unit, jobtitle, stop, cdg,

How is the *minorType* attribute related to *majorType*

Lookup Types

Inspect the words annotated with *Lookup* by clicking on the *Annotation List* tab.

List 10 values that are assigned to the *majorType* attribute!

facility_key, loc_key, address, number, time_modifier, date_unit, title, time, country_adj person_first, organization, location, currency_unit, jobtitle, stop, cdg,

How is the *minorType* attribute related to *majorType*

The minorType attribute specifies details about the type that is specified in the majorType.

ANNIE Components

ANNIE English Tokeniser

- Splits text into tokens
- Distinguishes between *Punctuation*, *SpaceToken*, *Word*, ...
- Adds additional attributes like
  - kind: word/number
  - orth: lowercase/uppercase
  - length
  - ...

ANNIE Components

ANNIE Gazetter

- Looks up words in predefined wordlists
- Adds *Lookup* annotations which often have a minorType and a majorType attribute
- Lookup lists are stored in text files which can easily be extended

ANNIE Components

ANNIE Sentence Splitter

- Identifies sentences and splits them
- Produces annotations *Sentence* and *Split*

ANNIE Components

ANNIE POS Tagger

- Assigns the *Token* annotations an attribute *category* that stores the part of speech (NN = singular or mass noun, . . .). There are 36 POS tags **https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html**

ANNIE Components

ANNIE NE Transducer

- Set of JAPE grammars that identify semantic units like name, date, address. . .

Regular Expressions

- Efficient string matching technique
- Search for character series or patterns of characters
- Written in a formal language (can differ between implementations)
- Can be represented as finite state automata

Regular Expressions

The *Kleene Star* *

Matches **zero or more** occurrences.
"$A*$" matches "", "A", "AAA", ...

The +

Matches **one or more** occurrences.
"$A+$" matches "A", "AAA", ...

The ?

Matches **zero or one** occurrence.
"$AB?A$" matches "AA", "ABA"

More special characters exist but these are the most important ones

JAPE – Java Annotation Patterns Engine

- Finite state transduction over annotations based on regex
- Specified Grammar

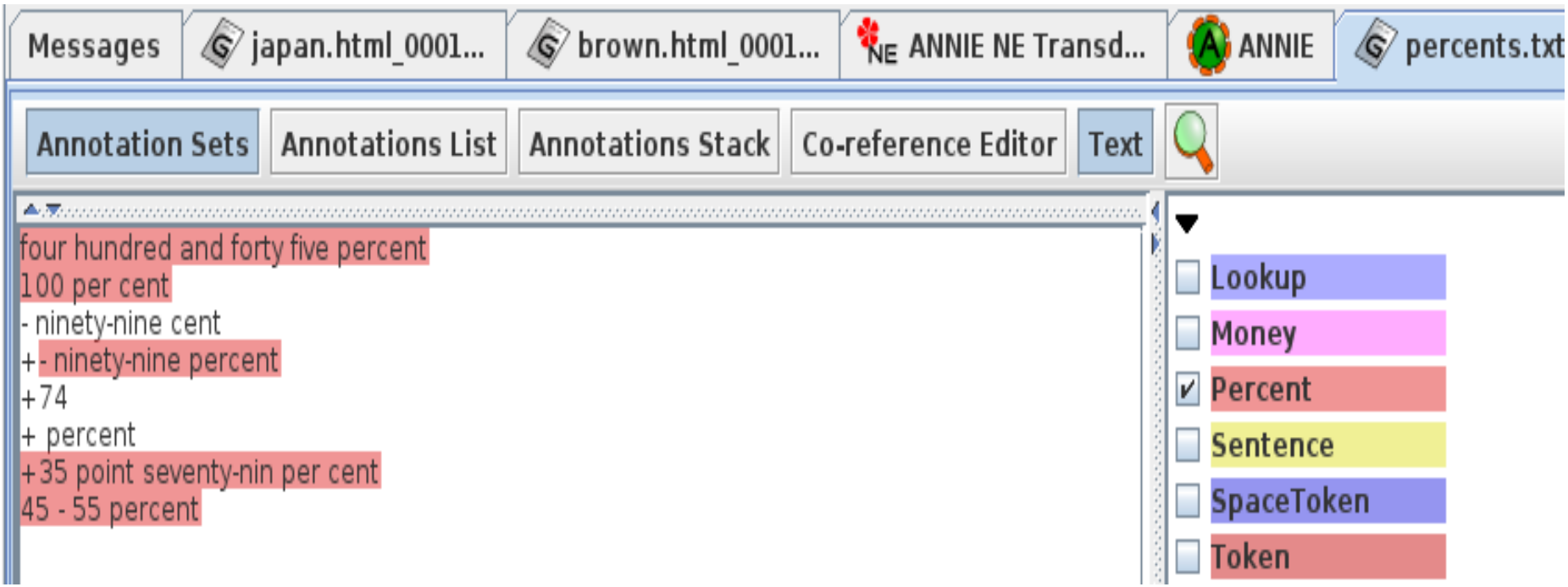
```
Phase: Number
Input: Token Lookup
Options: control = appelt

Rule: PercentBasic
// +20%
// minus 10 percent
// two point four percent
(({Token.string == " + "}|
{Token.string == "-"}|
{Token.string == "minus"})?
((AMOUNT_NUMBER|NUMBER_WORDS)
{Token.string == "point"}
)?
(AMOUNT_NUMBER|NUMBER_WORDS)
(PERCENT)
)
:number --- >
:number.Percent = {rule = "PercentBasic"}
```

JAPE

- A set of grammar rules that are applied in the same run is called a phase
- *Input* specifies the previous annotations a phase needs to access
- Macros store patterns for reusing them
- Rules have a Left Hand Side (LHS) to match a pattern and a Right Hand Side (RHS) to assign annotations
- Annotations can have attributes/features

1. four hundred and forty five percent
2. 100 per cent
3. $-$ ninety-nine cent
4. $+-$ ninety-nine percent
5. $+74$
6. $+$ percent
7. $+35$ point seventy-nine per cent
8. $45-55$ percent

1. four hundred and forty five percent
2. 100 per cent
3. − ninety-nine cent
4. +− ninety-nine percent
5. +74
6. + percent
7. +35 point seventy-nine per cent
8. 45 − 55 percent



More on JAPE rules

Why is Ban Ki-moon not recognized as Person annotation in the brown.html text? Why does the annotation creation work with Mirek Topolanek?

Before Mirek Topolanek there is a *Jobtitle* annotation. The word “secretary general” before Ban Ki-moon is not recognized as a title.
Solution: Either add general as a *Jobtitle* or add Ban Ki-moon to the gazetteer.

G20 JAPE Rule

```
Phase:  g20
Input:  Token
Options:  control = appelt

Rule:  G20
(
{Token.string == "G"}
{Token.string == "20"}
):g20countries
-->
:g20countries.VeryImportantCountries = {population =
"4385"}
```

Note: works up to v6.x!

WordNet (v3.0)

Find the correct WordNet sense number for the highlighted words in the following sentences!

> *In botany, a tree is a **plant** with an elongated **stem**, or **trunk**, supporting **leaves** or **branches**.*

1. plant#2
2. stem#2
3. trunk#1
4. leaf#1
5. branch#2

(⋆) Run, e.g.,: `wn leaf -over -s`

WordNet (v3.0)

Find the correct WordNet sense number for the highlighted words in the following sentences!

> *A manufacturing **plant** is an industrial site, usually consisting of buildings and machinery, or more commonly a **complex** having several buildings, where workers manufacture **good**s or operate machines processing one **product** into another.*

1. plant#1
2. complex#4 (why?)
3. good#4
4. product#2

(⋆) Run, e.g.,: `wn plant -over -s`

Finding Relations

WordNet defines different relations between two words:

1. synonym: X and Y share almost the same meaning
2. hyponym: X is a type of Y
3. hypernym: X is a generalization of Y
4. holonym: X consists (partially) of Y
5. meronym: X is part of Y

As you can see some relations just are the inverse of each other.

Finding Relations

Some examples:

- *mouse* is a hyponym of *rodent*
- *mouse* is a hypernym of *field mouse*.
- *field mouse* is a hyponym of *rodent*.
- *serpent* is a synonym of *snake* (and the other way around)
- *motor* is a meronym of *car*
- *car* is a holonym of *tire*

Note that the relation between *field mouse* and *rodent* is inherited. If *mouse* is a hyponym of *rodent* and *field mouse* is a hyponym of *mouse* then *field mouse* is a hyponym of *rodent* as well.

Finding Relations

Find the relations given above between the following words in WordNet. The relations may either be direct or inherited. Also provide the respective sense number for each word.

1. stem – trunk

Finding Relations

Find the relations given above between the following words in WordNet. The relations may either be direct or inherited. Also provide the respective sense number for each word.

1. stem – trunk : stem is a direct hypernym of trunk
2. trunk – tree

Finding Relations

Find the relations given above between the following words in WordNet. The relations may either be direct or inherited. Also provide the respective sense number for each word.

1. stem – trunk : stem is a direct hypernym of trunk
2. trunk – tree : trunk is a part meronym of tree
3. tree – branch

Finding Relations

Find the relations given above between the following words in WordNet. The relations may either be direct or inherited. Also provide the respective sense number for each word.

1. stem – trunk : stem is a direct hypernym of trunk
2. trunk – tree : trunk is a part meronym of tree
3. tree – branch : tree is an inherited part holonym of branch
4. manufacturing plant – plant

Finding Relations

Find the relations given above between the following words in WordNet. The relations may either be direct or inherited. Also provide the respective sense number for each word.

1. stem – trunk : stem is a direct hypernym of trunk
2. trunk – tree : trunk is a part meronym of tree
3. tree – branch : tree is an inherited part holonym of branch
4. manufacturing plant – plant : manufacturing plant is a direct hyponym of plant
5. building complex – plant

Finding Relations

Find the relations given above between the following words in WordNet. The relations may either be direct or inherited. Also provide the respective sense number for each word.

1. stem – trunk : stem is a direct hypernym of trunk
2. trunk – tree : trunk is a part meronym of tree
3. tree – branch : tree is an inherited part holonym of branch
4. manufacturing plant – plant : manufacturing plant is a direct hyponym of plant
5. building complex – plant : building complex is a direct hypernym of plant
6. tree – plant

Finding Relations

Find the relations given above between the following words in WordNet. The relations may either be direct or inherited. Also provide the respective sense number for each word.

1. stem – trunk : stem is a direct hypernym of trunk
2. trunk – tree : trunk is a part meronym of tree
3. tree – branch : tree is an inherited part holonym of branch
4. manufacturing plant – plant : manufacturing plant is a direct hyponym of plant
5. building complex – plant : building complex is a direct hypernym of plant
6. tree – plant : tree is an inherited hyponym of plant

($\star$) Run e.g.,: `wn trunk -hypen`


Thank You!