Tokenization Group

CLASP Workshop
NYU
Nov. 7, 2009

Starting Questions

A. What is a token?

B. To what degree are segments at the character level relevant to tokenization?

C. Which regularizations are part of tokenization and which are part of some “higher” level of annotation?

Basic conclusion

- Tokenization should be rule-based, simple process that does not depend on external resources
- Complex decisions belong at later stages
- Tokenization should preserve whatever is needed for later processing

What is a token?

- Depends on orthography of your language
- Smallest unit of token types for your languge
- Can we assume Unicode input?
- Assume standoff annotation

No splitting, remapping

- No character can be part of two tokens
- No character should be mapped to another in tokenization step
- Each character is part of a token, but--
- Ordinary space is not a token for English, but
  - Location of spaces is significant
  - Stop-start info must be retained as features
- Types of whitespace are significant (line breaks, tabs)

All Higher Level Processing

- Normalization of contracted forms
- Correction of misspelled words
- Normalization of alternative spellings
- Aliases for named entities
- Identification of immutable multiword units
- Morphological analysis (inflectional and/or derivational)

Special cases = special purpose grammars

- Hyphenation
- Numbers
- URLs
- Chemical names
- Etc.
- Publish metadata for later processors?

Chinese, Japanese?, Arabic?

- Very few restrictions from orthography, tokenization step is very simple
- Word boundary detection is like tokenization, but handled on higher level
- Depends on dictionary, linguistic analysis