

Enriching the analysis data of Inputlog with linguistic information

Lieve Macken & Eric Van Horenbeeck

LT³, Language and Translation Technology Team
University College Ghent

Faculty of Applied Economics
University of Antwerp



Collaborative project

- Funded by the Flanders Research Foundation (FWO)
- Faculty of Applied Economics
University of Antwerp
 - Inputlog
- Language and Translation Technology Team (LT³)
University College Ghent
- Work in progress
 - English and Dutch

LT³

- Research group embedded in the Faculty of Translation Studies
- Natural Language Processing / Computational Linguistics
- Focus on Language and Translation Technology

Enrich the analysis data of Inputlog

What kind of linguistic annotations?

- Shallow linguistic analysis
 - Part-of-speech, lemma, chunk information
- Word frequency information
- Syllable boundaries
- ...

Enrich the analysis data of Inputlog

Aim

- Valuable basis for more linguistically-oriented process research
- Enables analysis at a higher level, e.g.
 - Analysis of pauses
 - Do they occur at linguistic boundaries?
 - Lexical replacements during revision
 - Belong replaced words to the same word class?
 - Are higher frequency words replaced by lower frequency words?

Enrich the analysis data of Inputlog

Challenges

- 1 Keystroke logging tools log at letter level ⇔ NLP tools work with sentences and words
→ Aggregate the logged process data from the letter level (keystroke) to the word level
- 2 Keystroke logging tools log process data ⇔ NLP tools are designed for clean and grammatically correct text
→ Treat final product, deletions and spelling corrections in a different way

Enrich the analysis data of Inputlog

Input = S-notation

- Parse S-notation
- Separate product and process data
 - Final writing product
 - Process data
 - Spelling corrections (corrections at the level of the word)
 - Deletions

Example S-notation



S-notation

Th[r₁]₁¹e q{u}²ick|₂ brown [dog]⁵{f⁵}[i]⁷{o}⁸|₈x⁶}]₇ jumps
over the [{old }]₃⁴]}₃lazy [d]₃³{fox}⁹]₁₀{dog}¹⁰]₁₁. The end[]₁₂¹²!

Example S-notation



S-notation

Th[r₁]₁¹e q{u}²ick|₂ brown [dog]⁵{f⁵}[i]⁷{o}⁸|₈x⁶}]₇ jumps
over the [{old }]₃⁴]}₃lazy [d]₃³{fox}⁹]₁₀{dog}¹⁰]₁₁. The end[]₁₂¹²!

S-notation without indices

Th[r]e q{u}ick brown [dog]{f}{i}{o}x jumps
over the [{old }]}lazy [d]{fox}{dog}. The end[.]!

Example S-notation



Tokenization and sentence splitting

- Split text in sequences of sentences
 - Sentence boundary after sentence-final punctuation mark followed by capital letter
 - No sentence boundary after full stop that is part of an abbreviation
- Split sentence into sequence of words
 - Strip off punctuation marks that are not part of an abbreviation

Th[r]e q{u}ick brown [dog]{f}{i}{o}x jumps
over the [{old }]}lazy [d]{fox}{dog}.
The end[.]!

Example S-notation



Final product

Delete everything in between square brackets

Th[r]e q{u}ick brown [dog]{f}{i}{o}x jumps
over the [{old }]}lazy [d]{fox}{dog}.
The end[.]!

Example S-notation



Deletions

Keep all words/phrases with only word-external square brackets

Th[r]e q{u}ick brown [dog]{f}{i}{o}x
jumps over the [{old }]}lazy [d]{fox}{dog}.
The end[.]!

Example S-notation



Spelling corrections

Keep all words with word-internal square or curly brackets

Th[r]₁]₁¹e q{u}²ick|₂ brown [dog]⁵{f⁵}[i]⁷{o}⁸|₈x⁶}]₇ jumps
over the [{old }]₃⁴]}₃lazy [d]₃³{fox}⁹]₁₀{dog}¹⁰]₁₁. The end[]₁₂¹²!

Th[r] ₁] ₁ ¹ e	Th[r] → The
q{u} ² ick ₂	qick → q{u}ick
[dog] ⁵ {f ⁵ }[i] ⁷ {o} ⁸ ₈ x ⁶ }] ₇	[dog] → {f}{ix} → f{i}x → f{o}x
[d] ₃ ³ {fox} ⁹] ₁₀ {dog} ¹⁰] ₁₁	[d] → [fox] → {dog}

Table Format



# Revisions	Index	Product	Deletions	Word level corrections
1	1	The		ʰɪ[ɹ]ʰe
1	2	quick		kw[ɪ]kw[ɪk]
4	5,6,7,8	fox		[fɔks]ʰ(r)ʰ(i)ʰ(o)ʰ(s)ʰ
		jumps		
		over		
		the		
1	11	lazy		[oɪ]ʰ
3	3,9,10	dog		[dɔ]ʰ[fɔks]ʰ(dɔg)ʰ
		.		
		The		
		end		
1	12	I		[ɪ]ʰ

Linguistic annotations



Shallow Linguistic analysis

- Part-of-speech tagging
- Lemmatization
- Chunking

Data extracted from S-notation

- Final product
- Deletions (in context)

Shallow Linguistic analysis



Part-of-speech tagging

- Aka grammatical tagging or word-category disambiguation
- Assign PoS code to each orthographic token
- Different PoS tag sets
 - English: Penn Treebank tag set
 - coarse-grained tag set
 - 45 distinct tags
 - e.g. VBZ = Verb, 3rd person singular present
 - Dutch: CGN tag set
 - fine-grained tag set: word class
 - wide range of morpho-syntactic features
 - 316 distinct tags
 - e.g. WW(pv,tgw,met-t) = Verb, finite verb, present tense, with -t

Shallow Linguistic analysis



Part-of-speech tagging

- Fully automatic process
- Taggers are trained on an annotated corpus
 - Learn word/tag frequencies, tag sequence probabilities, and/or rule sets
 - Training corpus: 1.5M tokens, 68K sentences
- Need surrounding local context to determine the proper tag
 - Typically a window of two to three words and/or tags
- Accuracy figures in the range of 97-98%
- Tool: **LeTSTAG**

Shallow Linguistic analysis



Lemmatization

- Determine the lemma (base form) for an orthographic token
 - Verbs → infinitive
 - walking → walk
 - Other word categories → stem (word form without affixes)
 - shoes → shoe, better → good
- Lemmatizer uses the disambiguated predicted PoS codes to disambiguate ambiguous word forms
 - We have a **meeting** tomorrow → meeting
 - We are **meeting** tomorrow → meet
- Trained on Celex
- Accuracy figures in the range of 97-99%
- Tool: **LeTsLEMM**

Shallow Linguistic analysis



Chunking

- Grouping of words into non-overlapping and non-recursive chunks
- Superficial sentence analysis
- Rule-based chunkers
 - Small set of constituency and distitency rules
 - Uses the disambiguated predicted PoS codes and lemmata to group the words
- IOB tags
 - Inside a chunk
 - Outside a chunk
 - Beginning of a chunk
- Accuracy figures in the range of 94-95%
- Tool: **LeTsCHUNK**

Shallow Linguistic analysis



[NP The_{DT} quick_{JJ} brown_{JJ} fox_{NN}] [VP jumps_{VBZ}]

B-NP I-NP I-NP I-NP B-VP

[PP over_{IN} [NP the_{DT} old_{JJ} lazy_{JJ} dog_{NN}]] .

B-PP B-NP I-NP I-NP I-NP O

Table Format



PRODUCT				DELETIONS			
Product	PoS	Lemma	Chunk info	Deletions	PoS	Lemma	Chunk info
The	DT	the	B-NP				
quick	JJ	quick	I-NP				
brown	JJ	brown	I-NP				
fox	NN	fox	I-NP				
jumps	VBZ	jump	B-VP				
over	IN	over	B-PP				
the	DT	the	B-NP				
				[old] ¹¹	JJ	old	I-NP
lazy	JJ	lazy	I-NP				
dog	NN	dog	I-NP				
.	.	.	O				
The	DT	the	B-NP				
end	NN	end	I-NP				
				[.] ¹²	.	.	O
!	.	!	O				

Linguistic annotations



Additional annotations

- Syllabification
- Word Frequency information

Syllabification



Syllabification

- Separation of a word into syllables
- Data-driven approach
 - Derive syllable boundaries from an evidence base of pre-syllabified words
 - Trained on Celex
- Examples:
 - En: sum-mer-school key-stroke log-ging
 - Nl: hot-ten-tot-ten-ten-ten-ten-toon-stel-ling
- Accuracy figures in the range of 92-95%
- Demo for English and Dutch available on <http://lt3.hogent.be/en/tools/timbl-syllabification/>

Word frequency



Types versus tokens

- Counting words in corpora
 - Types are word forms
 - Tokens are occurrences of word forms
 - The quick brown fox jumps over the old lazy dog
 - 10 tokens, 9 types
- Corpus size vs. vocabulary size
 - Corpus size = the number of tokens in a corpus
 - Vocabulary size = number of types in a corpus

Word frequency



How to define a type?

- Do uppercase and lowercase words belong to the same type?
 - I go to **New** York
 - **New** cars are ugly.
- Do identical word forms with different PoS belong to the same type?
 - How to **book** a cheap flight?
 - I read an interesting **book**
- Do morphological variants belong to the same type?
 - books, book
 - send, sending, sent, sends

Word frequency



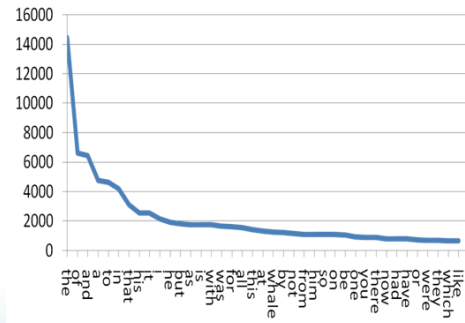
Word frequency distributions

- Word frequencies in corpora
 - Few high-frequent words
 - Many low-frequent words
- Zipf's law

"Given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table"

→ The most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, ...

Word frequency



Word frequencies in Moby Dick
Source: <http://searchengineland.com/the-long-tail-of-search-12198>

Representing word frequencies



Corpus

- Corpus of Contemporary American English (COCA)
 - 410 million word corpus
 - Frequency list available at <http://www.wordfrequency.info/>

Absolute frequencies: types and counts

the	22995878
and	11239776
logging	1855
multilateral	1855
keystroke	110
raindrop	110

Representing word frequencies



Frequency Ranks

- Word counts can be large numbers
 - Useful to present frequency information as ranks
- Assign a rank to frequency classes
 - rank 1 = most frequent word
 - rank 2 = second most frequent word
 - ...
- Types with the same absolute frequency are assigned the same a rank range
- Allow you to select the top *n* most frequent words

Representing word frequencies



Frequency ranks: types and ranks

the	1	1
and	2	2
logging	14196–14203	14203
multilateral	14196–14203	14203
keystroke	72093–72437	72437
raindrop	72093–72437	72437

Table Format



Product	Deletions	Syllabification	Absolute freq	Freq rank
The		the	22995878	1
quick		quick	30549	1378
brown		brown	22508	1907
fox		fox	5678	6394
Jumps		jumps	3714	8741
over		o-ver	314614	123
the		the	22995878	1
	[old] ¹¹	old	190027	185
lazy		la-zy	3438	9253
dog		dog	35880	1158
.				
The		the	22995878	1
end		end	129052	297
	[.] ¹²			
!				

Questions?

Suggestions?

