DISCOURSE LEXICONS FOR MULTIPLE LANGUAGES AND THEIR USE FOR GENDER PROFILING

Ben Verhoeven & Walter Daelemans

CLiPS Research Center, University of Antwerp, Belgium

Presented at CLIN 2018, Nijmegen

26/01/2018

IN A NUTSHELL

- Create discourse lexicons
- Multiple languages

- Methods from statistical MT
- Combined with Penn Discourse Treebank

- Gender profiling as application: English, Dutch, German
- Discourse information as features
- Comparison with knowledge-based lexicon and discourse parser

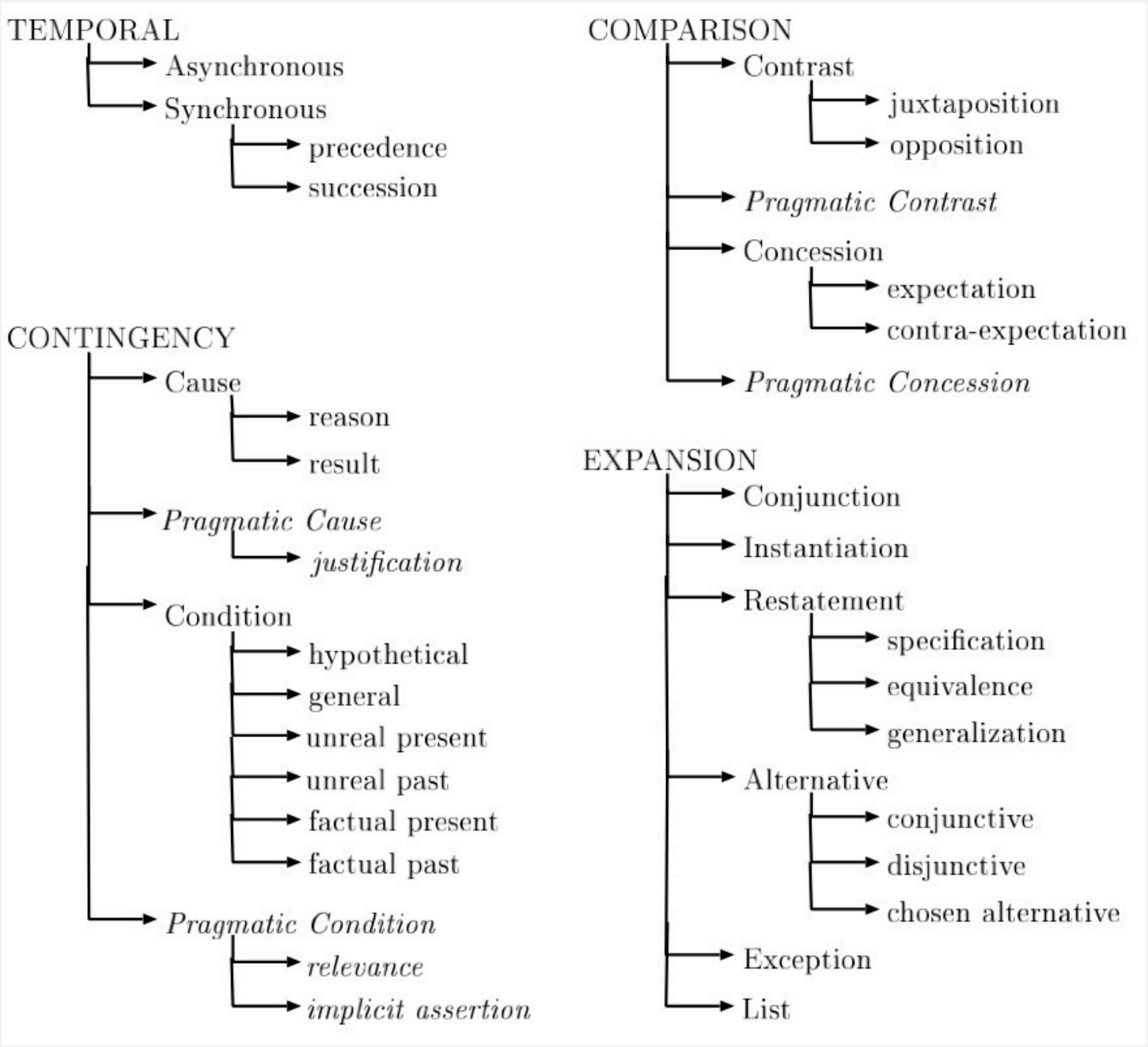
DISCOURSE ANALYSIS

- Discourse
  - Level of information in text above sentence level.
  - Concerned with structure and coherence
- Connectives
  - “in contrast” or “moreover”
- Models
  - Rhetorical structure theory (RST)
  - Penn Discourse Treebank (PDTB)

DISCOURSE ANALYSIS

- Discourse relations
  - Connection between two arguments
  - E.g. “COMPARISON”

- Limited number of practically usable tools
- Few languages, mainly English


TEMPORAL
Asynchronous
Synchronous
precedence
succession
CONTINGENCY
Cause
reason
result
Pragmatic Cause
justification
Condition
hypothetical
general
unreal present
unreal past
factual present
factual past
Pragmatic Condition
relevance
implicit assertion
COMPARISON
Contrast
juxtaposition
opposition
Pragmatic Contrast
Concession
expectation
contra-expectation
Pragmatic Concession
EXPANSION
Conjunction
Instantiation
Restatement
specification
equivalence
generalization
Alternative
conjunctive
disjunctive
chosen alternative
Exception
List

GENDER PROFILING

- Predict whether author of text is male or female

- Supervised machine learning
- Common features: word and character n-grams

DISCOURSE FEATURES FOR AUTHOR PROFILING

- Soler-Company & Wanner (2017)
  - RST discourse parser by Surdeanu et al. (2015)
  - Features:
    - Frequency of discourse relations
    - Depth and width of discourse tree

DIMUL: DISCOURSE FEATURES FOR MULTIPLE LANGUAGES

- Penn Discourse Treebank (PDTB)
  - English text corpus
  - Occurrences of discourse connectives annotated with discourse relation
  - 100 unique connectives
- Example
  - “moreover” only occurs with EXPANSION as associated relation (weight = 1.0)

DIMUL: DISCOURSE FEATURES FOR MULTIPLE LANGUAGES

- Extrapolate to other languages
  - Phrase tables from statistical MT
    - Words/phrases that are each other's best translations
    - Weights for collocation strength
  - Lopes et al. (2015) used Europarl
- Example
  - "moreover" has 5 Dutch translations: *trouwens, verder, voorts, bovendien de, bovendien in*
  - All 5 receive relations of source word: EXPANSION (weight 1.0)

DIMUL: DISCOURSE FEATURES FOR MULTIPLE LANGUAGES

- Example
  - But *trouwens* has a second source word: "besides"
  - "besides" in PDTB
    - EXPANSION (weight 0.9474)
    - COMPARISON (weight 0.0526)
  - Weight them by collocation strength (product and rescale)
    - besides-trouwens: 0.0037
    - moreover-trouwens: 0.1068

DIMUL: DISCOURSE FEATURES FOR MULTIPLE LANGUAGES

- Example
  - But *trouwens* has a second source word: "besides"
  - Relations are weighted and rescaled into the final target relations for *trouwens*
    - EXPANSION (weight 0.9982)
    - COMPARISON (weight 0.0018)
- Also experiments using the connectives in the lexicon, not their relations

MODEL COMPARISONS

- DimLex
  - Knowledge-based manually crafted discourse lexicon for German (Stede, 2012)
  - Features: counts of discourse relations and connectives
- RST parser
  - RST discourse parser for English by Surdeanu et al. (2015)
  - Features: counts of the discourse relations normalized by the document length
- Function words
  - Lists of function words compiled from different resources
  - Features: function word counts

DATASETS

Corpus	Language	Genre	Female	Male	Total
New York Times (NYT)	English	News	155,219	153,916	309,135
Het Laatste Nieuws (HLN)	Dutch	News	72,161	72,163	144,324
Blogger-NL	Dutch	Blogs	84,363	84,363	168,726
Blogger-DE	German	Blogs	238,105	238,105	476,210
Blog Authorship	English	Blogs	308,358	308,358	616,716

EXPERIMENTS

- Supervised machine learning
- Two classes: male and female
- 10-fold cross validation
- Baseline: 0.50

FEATURE ANALYSIS

- Genre-related: JUSTIFICATION is strongly male in news, strongly female in blogs.
- Language-related:

	English	Dutch	German
Female		*concession*	*concession*
	pragmatic contrast	opposition	*unreal past*
	implicit assertion		*unreal present*
Male	unreal present	**unreal present**	
	temporal		**temporal**
	contingency		
	factual present		
	unreal past		
	concession		

DISCUSSION

- Discourse of a text contributes to gender prediction
  - Much better on blogs than news
    - DiMuL: 0.59-0.64 vs. 0.51-0.54
    - N-grams: 0.85-0.90 vs. 0.62-0.68
  - English blog authorship corpus is an outlier (more like news)
  - Logistic Regression overall best
- Feature types
  - The more features, the better the system
  - dimulcat1 < dimulcat2 < dimulcat3 < dimulconn < funcwords

DISCUSSION

- Two important comparisons
  - dimulcat3 > dimlexcat3 ($p < 0.001$)
    - 0.62 vs. 0.60
  - rst2 > dimulcat3 ($p < 0.001$)
    - 0.65 vs. 0.54

THANK YOU

- @verhoevenben
- ben.verhoeven@uantwerpen.be