Analyzing linguistic variation: From corpus query towards feature discovery

Elke Teich

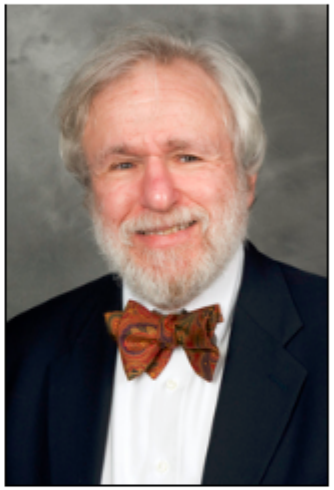
Saarbrücken, Germany

Collaborators

Marilisa Amoia, Stefania Degaetano-Ortlieb, Hannah Kermes,
Jörg Knappen, Ekaterina Lapshinova-Koltunski ,
Richard Litauer, José Manuel Martinez Martinez,
Mihaela Vela, Katja Weis
(Universität des Saarlandes, Saarbrücken)
&
Peter Fankhauser
(Institut für Deutsche Sprache, Mannheim)

Deutsche Forschungsgemeinschaft DFG
M2CI CLUSTER OF EXCELLENCE
UNIVERSITÄT DES SAARLANDES
CLARIN-D
Bundesministerium für Bildung und Forschung

Eugene Charniak



University Professor of Computer Science

Contact Information

Box 1910
Brown University
Providence, RI 02912
Email: ec at cs.brown.edu
Personal home page:
http://www.cs.brown.edu/~ec/

Research Areas

Artificial Intelligence
Natural Language Processing
Machine Learning

Courses Taught

CSCI0020 The Digital World
CSCI2410 Statistical Models in Natural-Language Understanding

Research Interests

Eugene Charniak is interested in programming computers to understand language so that they will be able to perform such tasks as answering questions and holding a conversation. This is far beyond our current capabilities, so research proceeds by dividing the problem up into manageable subparts. Prof. Charniak's research is called "statistical language learning." He and his students write programs that collect statistical information about language from large amounts of text, then apply the statistics to new examples. For example, much of his recent research has been on statistical models of syntactic parsing—grammatically identifying parts of speech and learning the rules for sentence formation, an exercise akin to the sentence diagramming that

United States Patent [19]

Kucera et al.

[11] **Patent Number: 4,868,750**

[45] **Date of Patent: Sep. 19, 1989**

[54] **COLLOCATIONAL GRAMMAR SYSTEM**

[75] Inventors: **Henry Kucera**, Providence, R.I.; **Alwin B. Carus**, Newton, Mass.; **Jeffrey G. Hopkins**, Pawtucket, R.I.

[73] Assignee: **Houghton Mifflin Company**, Boston, Mass.

[21] Appl. No.: **106,127**

[22] Filed: **Oct. 7, 1987**

[51] **Int. Cl.**[4] **G06F 15/02**

[52] **U.S. Cl.** **364/419**; 364/900

[58] **Field of Search** ..... 364/419, 900 MS File, 364/200 MS File

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,661,924	4/1987	Okamoto et al.	364/900
4,703,425	10/1987	Muraki	364/419
4,724,523	2/1988	Kucera	364/419
4,742,481	5/1988	Yoshimura	364/419
4,747,053	5/1988	Yoshimura	364/419
4,750,122	6/1988	Kaji	364/419
4,760,528	7/1988	Levin	364/419
4,773,009	9/1988	Kucera	364/419

OTHER PUBLICATIONS

Choice of Grammatical Word-Class Without Global Syntactic Analysis: Tagging Words in the Lob Corpus, Marshall, Ian in *Computers and the Humanities* 17 (1983) 139–150.

Primary Examiner—Michael R. Fleming
Attorney, Agent, or Firm—Lahive & Cockfield

[57] **ABSTRACT**

A system for the grammatical annotation of natural language receives natural language text and annotates each word with a set of tags indicative of its possible grammatical or syntactic uses. An empirical probability of collocation function defined on pairs of tags is iteratively extended to a selected set of tag sequences of increasing length so as to select a most probable tag for each word of a sequence of ambiguously-tagged words. For listed pairs of commonly confused words a substitute calculation reveals erroneous use of the wrong word. For words with tags having abnormally low frequency of occurrence, a stored table of reduced probability factors corrects the calculation. Once the text words have been annotated with their most probable tags, the tagged text is parsed by a parser which successively applies phrasal, predicate and clausal analysis to build higher structures from the disambiguated tag strings. A voice/text translator including such a tag annotator resolves sound or spelling ambiguity of words by their differing tags. A database retrieval system, such as a spelling checker, includes a tag annotator to identify desired data by syntactic features.

17 Claims, 13 Drawing Sheets

```
<?xml version="1.0" encoding="UTF-8"?>
<metadata>
  <identifier>BrownCorpus</identifier>
  <title>Brown Corpus</title>
  <creator>Nelson Francis and Henry Kucera</creator>
  <mediatype>texts</mediatype>
  <collection>LanguageCommons</collection>
  <collection>data</collection>
  <description>A Standard Corpus of Present-Day Edited American English, for use with Digital Computers.
By W. N. Francis and H. Kucera (1964), Department of Linguistics, Brown University, Providence, Rhode Island, USA.
Revised 1971, Revised and Amplified 1979.</description>
  <date>1964</date>
  <year>1964</year>
  <subject>en</subject>
  <licenseurl>http://creativecommons.org/licenses/by-nc/3.0/</licenseurl>
  <pick>1</pick>
  <public>1</public>
  <publicdate>2010-07-15 08:00:12</publicdate>
  <addeddate>2010-07-15 07:56:54</addeddate>
  <uploader>stevenbird1@gmail.com</uploader>
  <updater>Steven Bird</updater>
  <updater>Steven Bird</updater>
  <updatedate>2010-07-15 08:09:43</updatedate>
  <updatedate>2010-07-16 19:10:58</updatedate>
  <language>en</language>
  <coverage>US</coverage>
  <updatedate>2010-09-06 04:32:50</updatedate>
  <updater>Steven Bird</updater>
</metadata>
```

W. N. Francis & H. Kucera, 1964, A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Department of Linguistics, Brown University, Providence, RI, USA

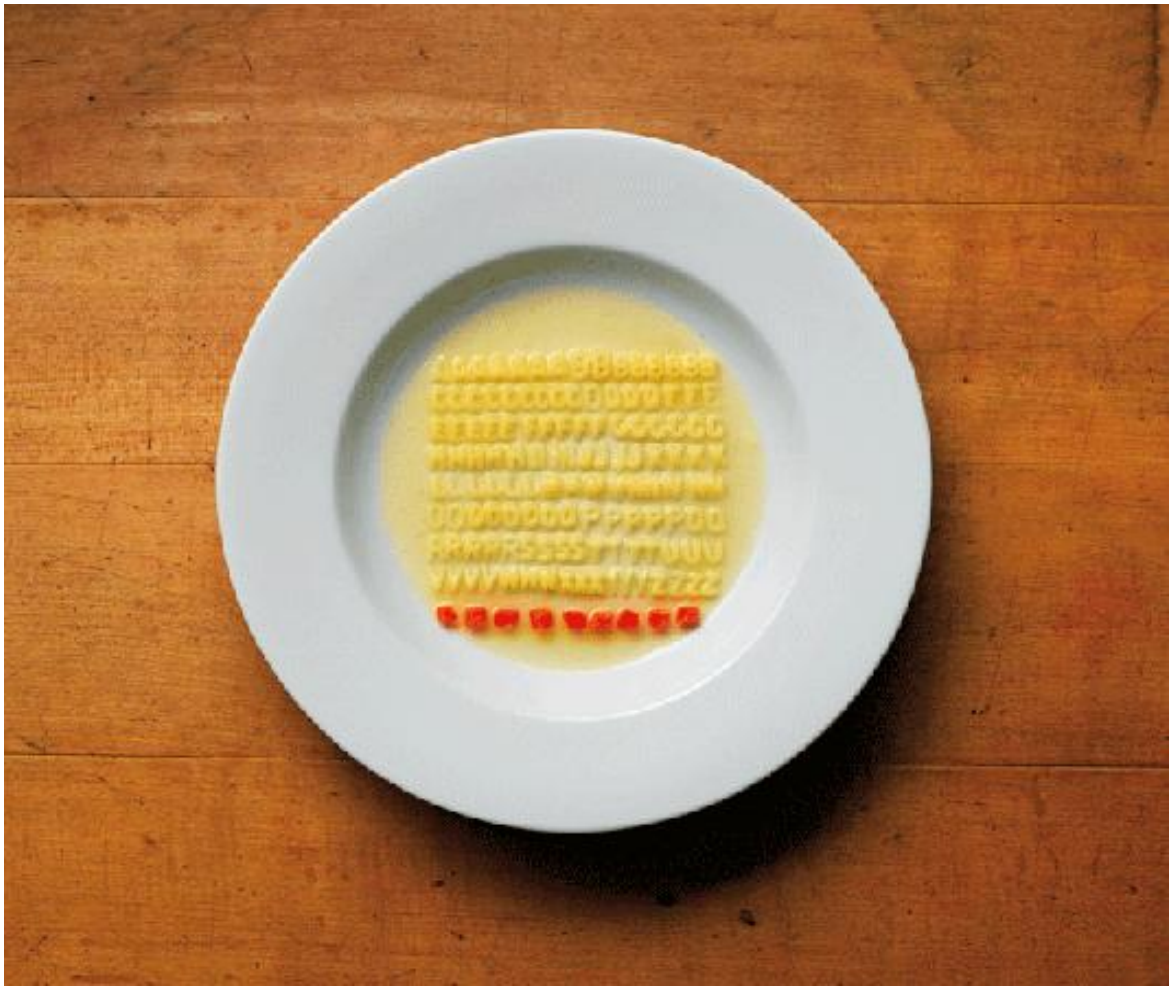
```
-u(Unix)--  BrownCorpus_meta.xml    All L1      (XML)----------
Loading regexp-opt...done
```

In a Station of the Metro

The apparition of these faces
in the crowd ;
Petals on a wet, black bough.

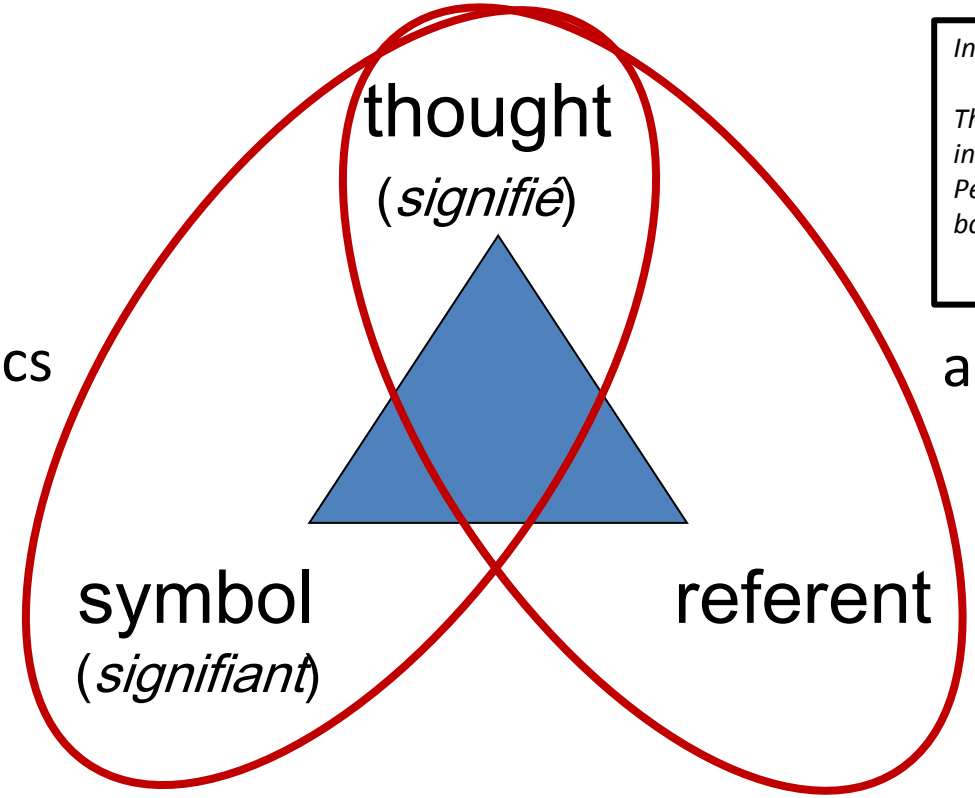
A linguist´s perspective

A linguist´s perspective



Urs Wehrli, *Kunst aufräumen*

linguistics



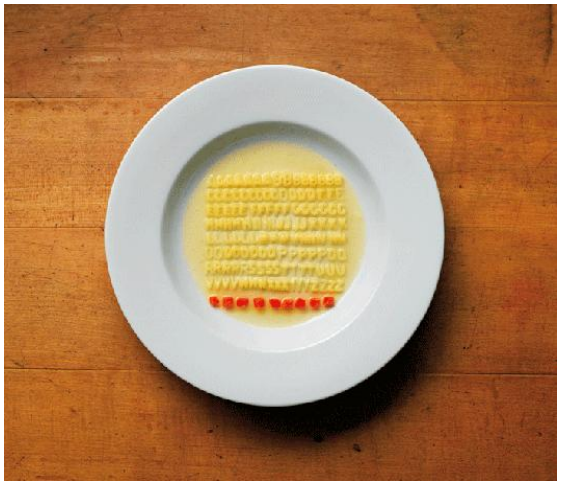
In a Station of the Metro

The apparition of these faces
in the crowd ;
Petals on a wet, black
bough.

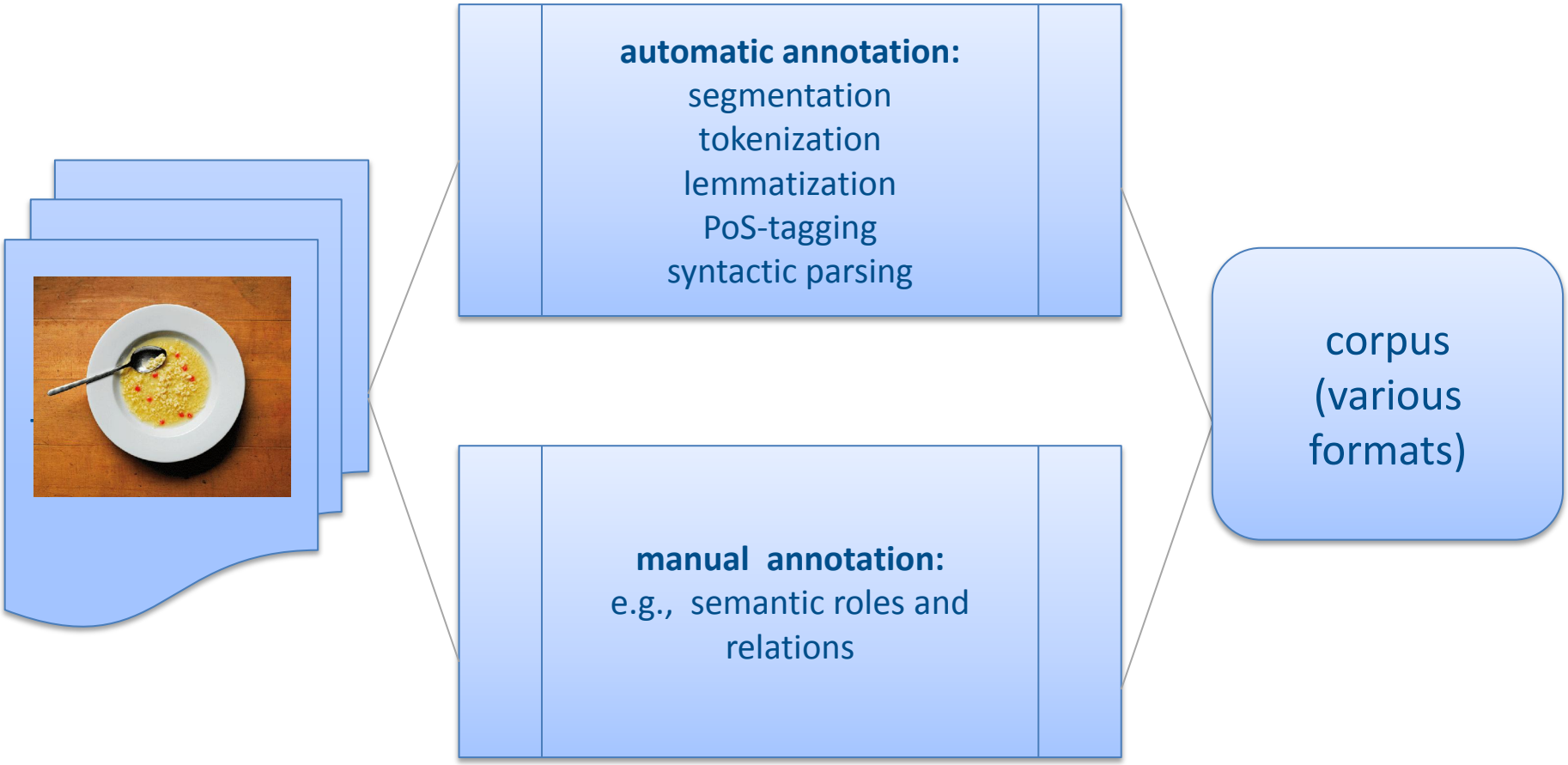
arts & humanities

(Odgen & Richards, 1923, *The meaning of meaning*)

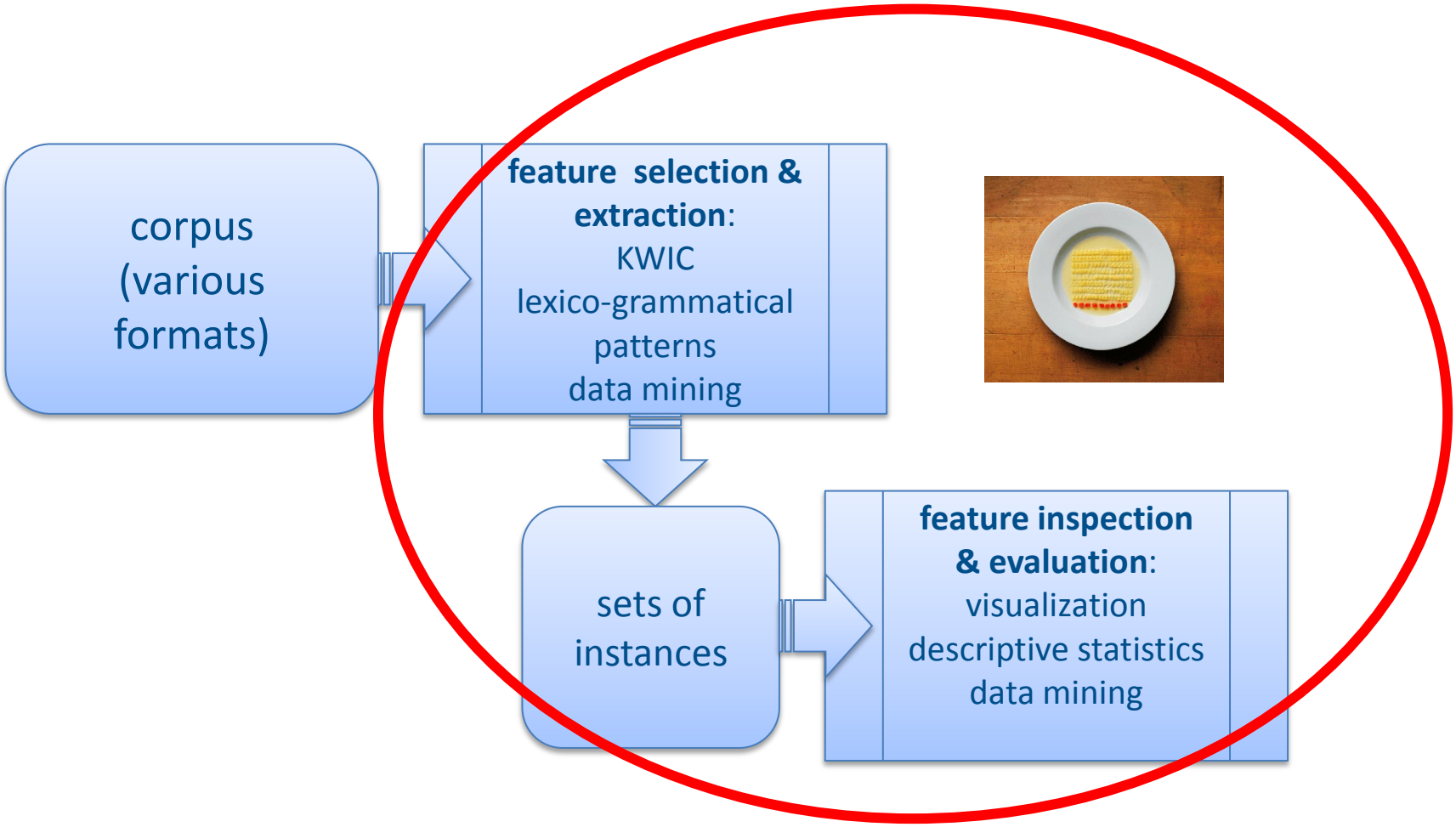
???



Processing Pipeline (1)



Processing Pipeline (2)



Analysis of linguistic variation

- Dialect/sociolect: regional/social variation
- Register: functional (situational) variation
- Register theory (Halliday, Biber a.o.)

> "A register is a cluster of associated features having a greater-than-random (or rather, greater than predicted by their unconditioned probabilities) tendency to co-occur."
>
> (Halliday, 1988:162)

- Registers are relatively stable in time; registerial repertoire of a language changes over time

Registers in Contact (RegiCo): Research Questions

New research fields are continuously developing (e.g., bioinformatics, mechatronics etc), often through contact between two disciplines (e.g., computer science – biology)

- What a … "contact discipl…
- How d… ir "seed discipl…

  → **sim…**
- Do they develop their own "language"?

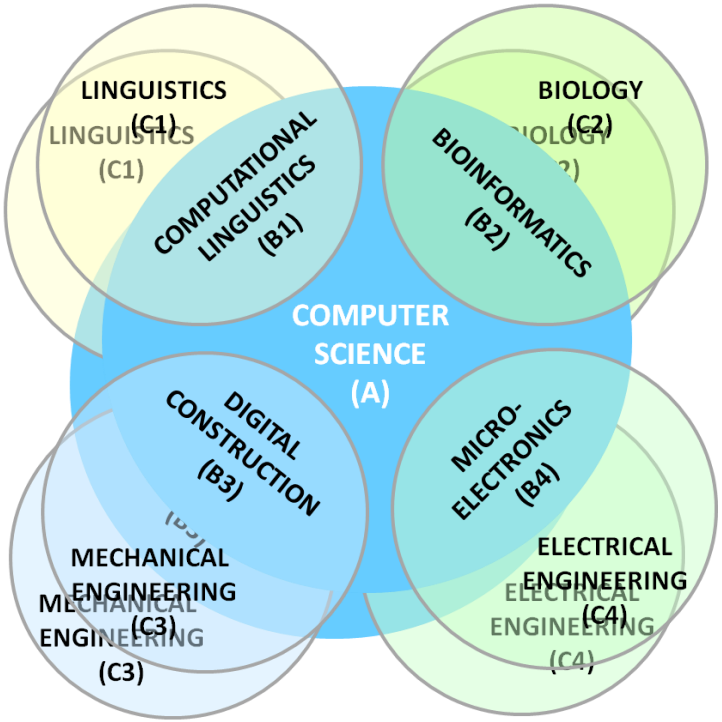
  → **distinctiveness**

What kind of resource is needed?

RegiCo: Corpus

English Scientific Text Corpus (SciTex)

- full English journal articles
- nine disciplines (register)
- two time slices (time):
  - DaSciTex (2000s)
  - SaSciTex (1970s/80s)
- approx. 34 million words



(Teich & Holtz 2009,
Teich & Fankhauser 2010,
Degaetano-Ortlieb et al. forthcoming)

RegiCo: Corpus encoding

- source: pdf
- formats
  - plain text
  - html
  - xml
  - CQP (Corpus Query Processor; Schmied, 1998; Evert, 2005)
- types of information
  - bib data: author, title, journal, year
  - discipline
  - logical structure (section, paragraph etc)
  - linguistic units: sentence, token
  - linguistic categories: lemma, part-of-speech (syntactic phrases)

RegiCo: Corpus encoding (CQP)

corpus position	token	lemma	PoS
0	This	this	DT
1			NN
2			VBZ
3			IN
4			PRP
5			VBP
6			DT
7	exact	exact	JJ
8	value	value	NN
9	.	.	PUN

Which analysis methods are needed?

+ text and sentence ids

RegiCo: Methods

- Compare acc. to:
  - register (r)
  - time (t)
- Compare in terms of:
  - lexico-grammatical feature (f1, f2, f3… fn) in a context (r, t)
- Contrast: relative similarity/difference (probability)

  unconditioned vs. conditioned probability, e.g.,

  p (f1) *vs.* p (f1 | r1)

  conditioned probabilities, e.g.,

  p (f1 | r1) *vs.* p (f1 | r2)

→ probability distance measures, e.g., statistical tests, clustering, classification

Analysis example (1)

- Stance/evaluation in scientific writing
- Indicators - examples

  Our algorithm is obviously a 2-approximation for the problem.

  It is obvious that dynamic backcalculation analysis is more advantageous than the static approach.

  Interestingly, these protocols invariably require the use of supersingular curves.

  It is interesting that the rates of lexicon growth are roughly similar to each other regardless of the algorithm used [...].

- Question: Are there differences across registers?

→ extraction, distribution, statistical testing

<s> []{0,3} "it|It" [pos="VB.*"][]{0,3} "important" @"that|to" within s;

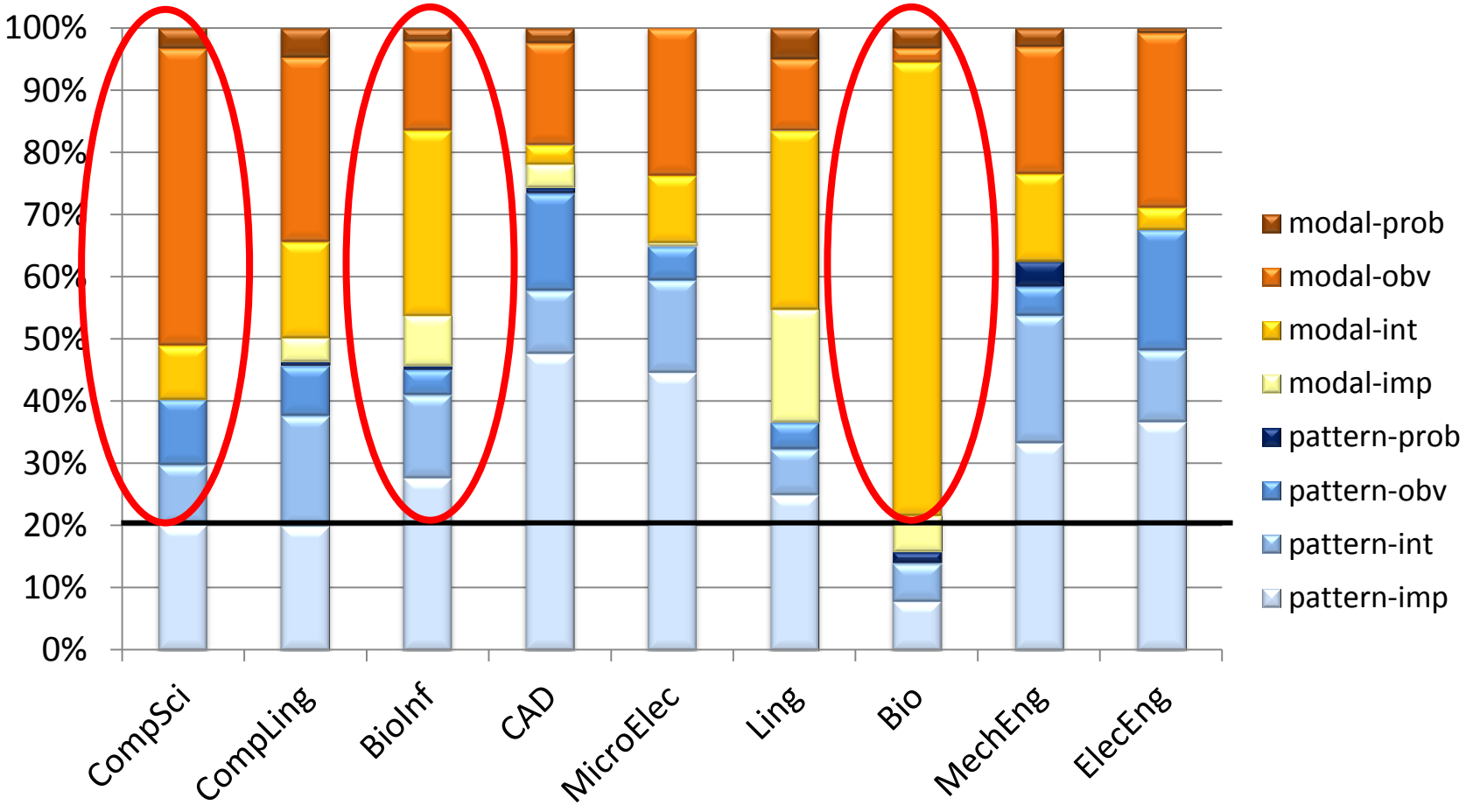
<It is important that> this work be extended to freely bubbling conditions where endogeneous bubbles interact with exogeneous ones .

<It is important that> this quantity be computed causally by a filter as s goes from 0 to T .

<It is therefore important to> account for the frictional stresses in the model.

<It was important to> adapt the recursion of forward and backward algorithm to the extended architecture of the HMMs .

	A	A	B1	B1	B2	B2	B3	B3	B4	B4	C1	C1	C2	C2	C3	C3	C4	C4	total
pattern-imp	56	3,44	57	3,50	76	4,67	78	4,79	82	5,04	71	4,36	45	2,76	75	4,61	73	4,48	613
pattern-int	27	1,66	40	2,46	34	2,09	16	0,98	27	1,66	17	1,04	35	2,15	38	2,33	22	1,35	256
pattern-obv	33	2,03	19	1,17	10	0,61	28	1,72	13	0,80	15	0,92	1	0,06	16	0,98	39	2,39	174
pattern-prob	0	0,00	3	0,18	2	0,12	1	0,06	1	0,06	0	0,00	12	0,74	7	0,43	0	0,00	26
modal-imp	6	0,37	29	1,78	49	3,01	25	1,54	24	1,47	81	4,97	44	2,70	7	0,43	8	0,49	273
modal-int	24	1,47	37	2,27	66	4,05	5	0,31	16	0,98	73	4,48	359	22,04	26	1,60	9	0,55	615
modal-obv	185	11,36	113	6,94	51	3,13	58	3,56	47	2,89	100	6,14	25	1,54	76	4,67	81	4,97	736
modal-prob	38	2,33	137	8,41	88	5,40	36	2,21	10	0,61	146	8,97	335	20,57	112	6,88	22	1,35	924



(Degaetano-Ortlieb & Teich, submitted)

Analysis example (2)

- Self-construal of actors in a scientific community
- Indicator: *we* + VERB

→ extraction, distribution, text classification (SVM)

- some results - what *we* do in
  - Computer Science (A): *prove, show, obtain* (´formal´)
  - Computational Linguistics (B1): *examine, implement, use* (´experimental´)
  - Linguistics (C1): *propose, suggest, argue* (´semiotic´), *feel, see* (´cognitive/emotive´)

(Teich & Fankhauser, 2010)

P: predicted class

C: class

C/P	A	B1	C1
A	210	**21**	**3**
B1	**19**	168	**47**
C1	**3**	**32**	199

- most misclassifications for Computational Linguistics (B1: 19+47+21+32), only few (3+3) between Computer Science (A) and Linguistics (C1)

  → **Computational Linguistics "in between"**

- Computational Linguistics more often misclassified as Linguistics (47) than as Computer Science (19), Linguistics more often misclassified as Computational Linguistics (32) than Computer Science misclassified as Computational Linguistics (21)

  → **Computational Linguistics closer to Linguistics**

Analysis example (3)

- Scientific writing is *technical, abstract* and *dense*
- Indicators:
  - technicality: *low type-token ratio*
  - abstractness: *many nouns*
  - density: *lexical density*
- Question: How typical are these features of scientific writing?

  Compare to non-scientific language (Brown/LOB)

→ extraction, distribution, text classification

(Teich & Fankhauser, 2010)

	DaSciTex	FLOB’	t-test	SVM
standardized TTR	34.0	45.3	29.5	97%
ADV	0.034	0.060	23.8	
N	0.33	0.27	-19.0	
lexical density	8.39	5.76	-18.4	
V	0.097	0.12	12.2	

single features: t-test; set of all features: SVM classifier

C/P	DaSciTex	FLOB’
DaSciTex	178	8
FLOB’	9	288

C: class; P: class predicted by SVM Classifier

Analysis example (4)

- formulaic expressions
- N-grams (4-grams), e.g.,

  the size of the, the fact that the (NP-based)

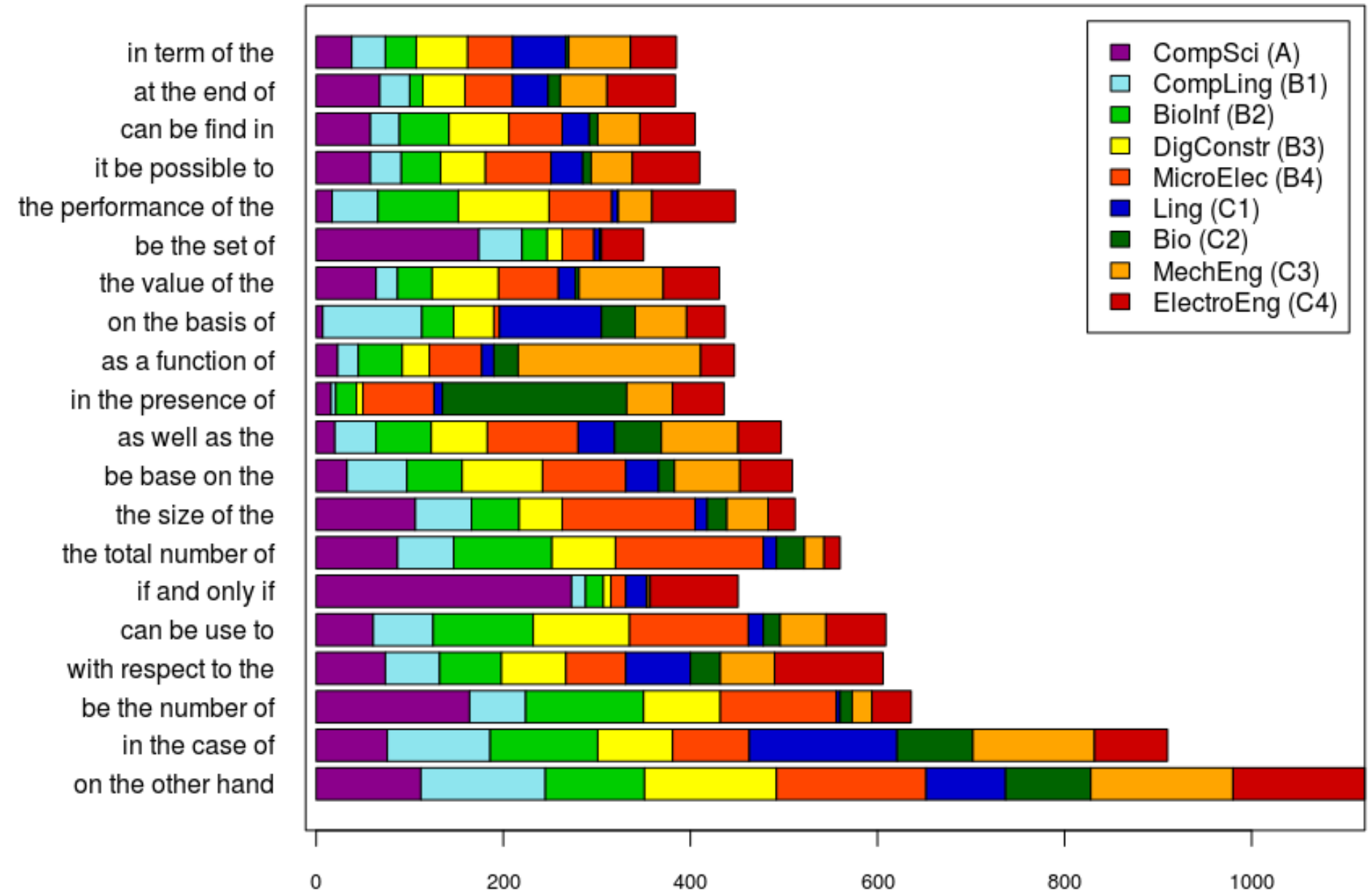
  with respect to the, in the case of (PP-based)

  can be used to, shown in table X (VP-based)
- Questions: Which ones do we find? Are there preferences acc. to register and document structure? What are their functions?

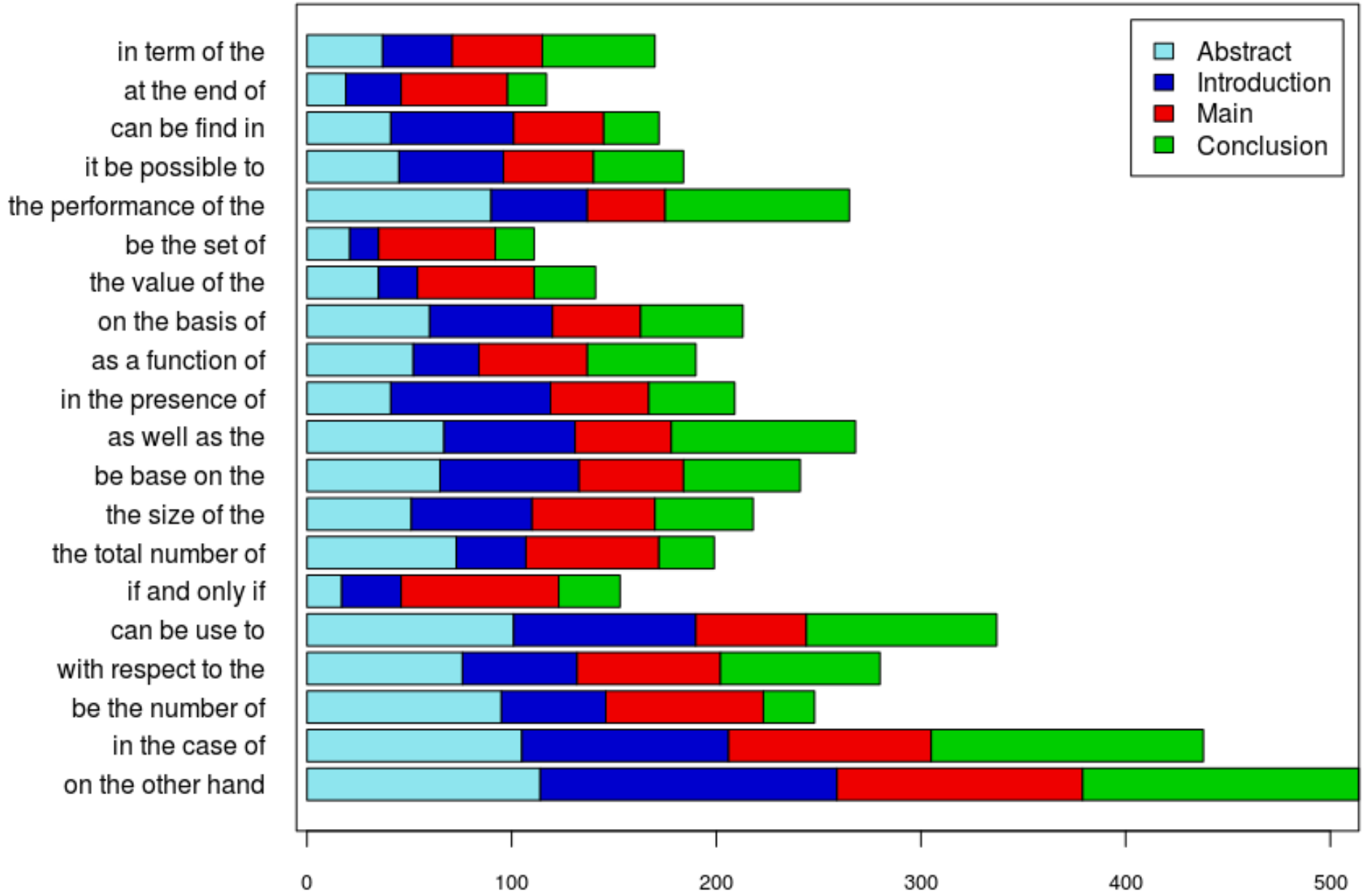
→ extraction, distribution, clustering

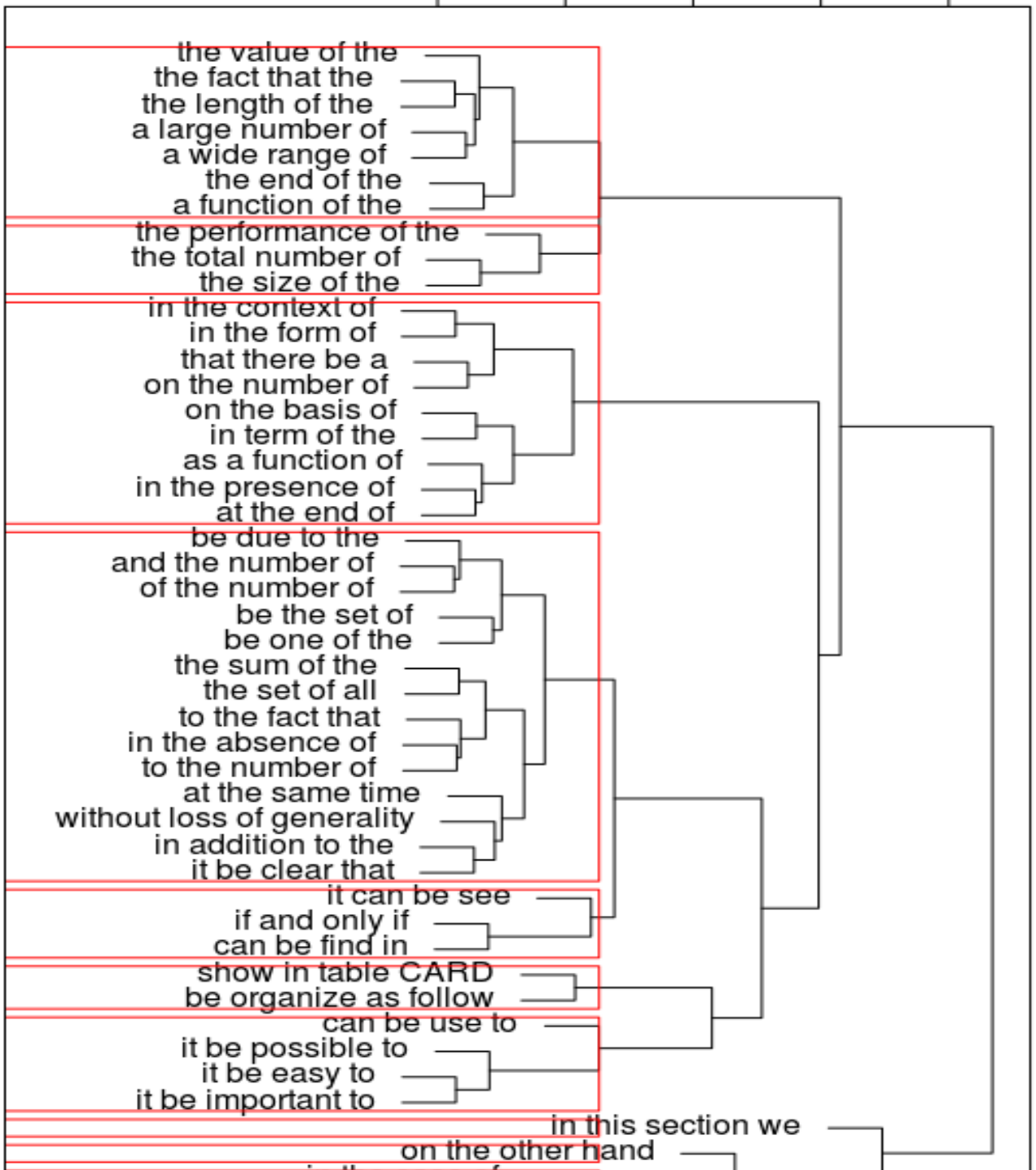
(Kermes, 2012; Kermes & Teich, in preparation)

Distribution: registers



Distribution: document structure




the value of the
the fact that the
the length of the
a large number of
a wide range of
the end of the
a function of the
the performance of the
the total number of
the size of the
in the context of
in the form of
that there be a
on the number of
on the basis of
in term of the
as a function of
in the presence of
at the end of
be due to the
and the number of
of the number of
be the set of
be one of the
the sum of the
the set of all
to the fact that
in the absence of
to the number of
at the same time
without loss of generality
in addition to the
it be clear that
it can be see
if and only if
can be find in
show in table CARD
be organize as follow
can be use to
it be possible to
it be easy to
it be important to
in this section we
on the other hand

Functions in discourse

4gram	clus	SimpsonEllis	Biber
on the other hand	1	RefExContComp	DiscTopElab
in the case of	2	RefExIntFram	RefExIntFram
be the number of	2	RefExQuant	
with respect to the	2	RefExIntFram	
be base on the	2	RefExIntFram	
as well as the	2	DiscMark	DiscTopElab
can be use to	3	StanceAbil	StanceAbil
it be possible to	3	StanceAbil	StanceAbil
it be easy to	3		
it be important to	3	StanceEval	StanceObl
if and only if	4		
can be find in	4	StanceAbil	
it can be see	4	RefExIdentFoc	
the total number of	5	RefExQuant	
the size of the	5	RefExTangFram	RefExTangFram
the performance of the	5		
as a function of	6	RefExIntFram	
in the presence of	6	RefExIntFram	RefExIntFram
on the basis of	6	RefExIntFram	RefExIntFram
in term of the	6	RefExIntFram	RefExIntFram
at the end of	6	RefExDeicLoc	RefExMulti
in the context of	6	RefExIntFram	
that there be a	6	RefExIdentFoc	
in the form of	6	RefExIntFram	RefExTangFram
on the number of	6		

feature selection & extraction:
KWIC
lexico-grammatical patterns
data mining

sets of instances

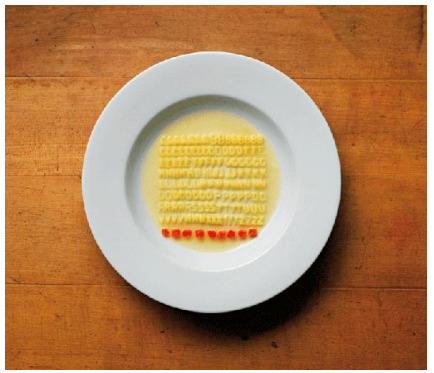
feature inspection & evaluation:
visualization
descriptive statistics
data mining

automatic annotation:
segmentation
tokenization
lemmatization
PoS-tagging

text collection

corpus

manual annotation:
e.g., semantic roles and relations



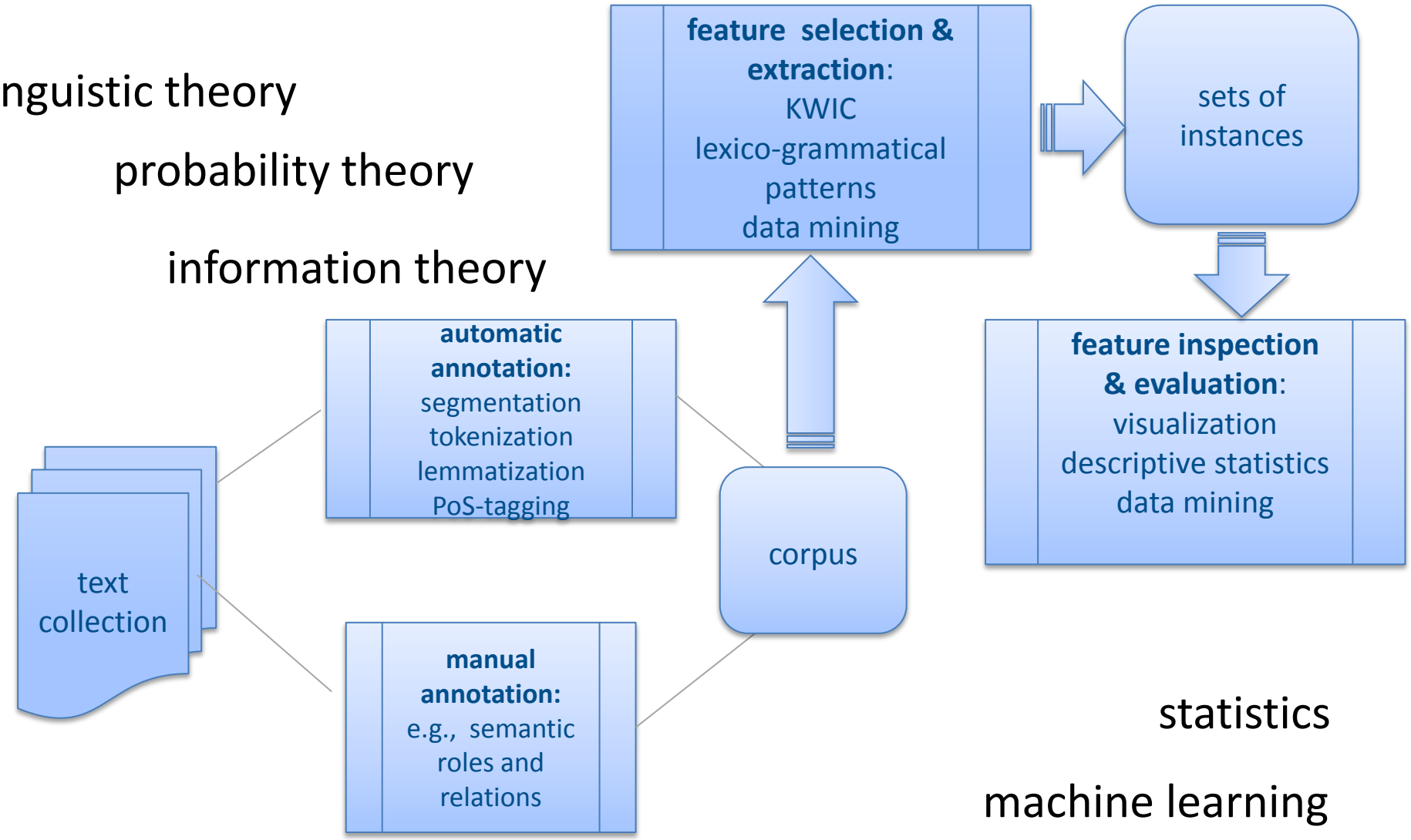
Modeling

formal grammar

linguistic theory

probability theory

information theory



statistics

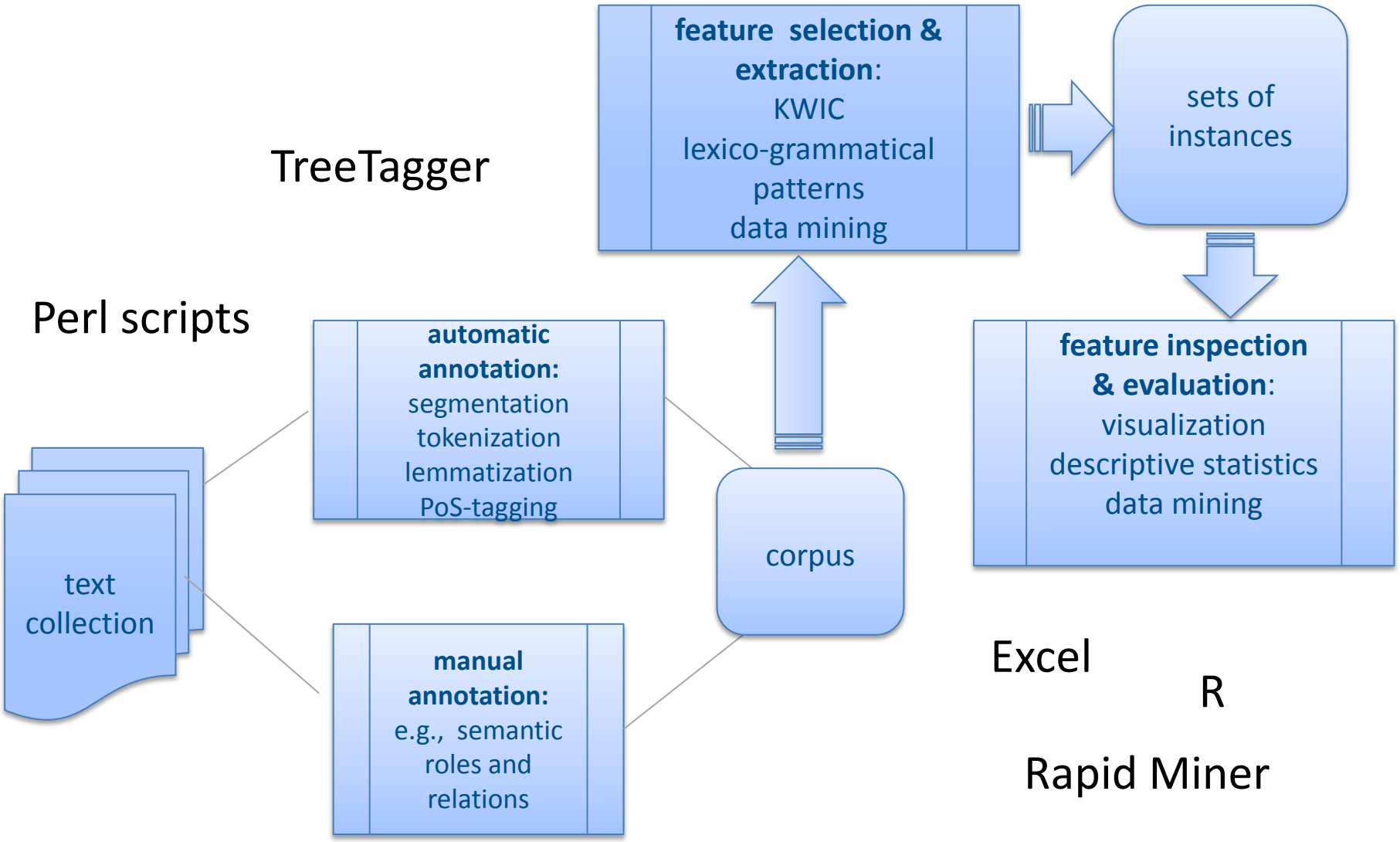
machine learning

Tools

CQP

TreeTagger

Perl scripts



Excel

R

Rapid Miner

Workflow model?

Requirements on ``data modeling'' for linguistics

- Resource: Corpus
  - flexibility: creation of new versions of a corpus ´´on the fly´´
  - accessibility: easy query/mining of a corpus
  - adressability: identify relevant objects of study

  → **agreement on object (unit) to be described**
- Computational processing : task-specific models rather than one overarching model

  → **each model can be tested for adequacy**

- Corpus Analysis: Tools
  - 1990‘s: build the ideal corpus tool (e.g., Mate, Nite projects)
  - 2000‘s: ideal corpus tool not possible; instead:
    - task-specific tools (e.g., pos-tagging)
    - compatibility among tools with different tasks (easy import/export, e.g., CQP → R, WEKA, Rapid Miner etc)
  - 2010‘s: frameworks for building processing pipelines, e.g., WebLicht (Clarin-D) XML-based TCP format

→ recognition of diversity in classification of object (& use your favorite tool)

``Eugene Charniak is interested in programming computers to understand language so that they will be able to perform such tasks as answering questions and holding a conversation. This is far beyond our current capabilities, so research proceeds by dividing the problem up into manageable subparts.´´
(http://www.cs.brown.edu/people/faculty/ec.html)