Degrees of Orality in Speech-like Corpora:

Comparative Annotation of Chat and E-mail

Eckhard Bick

University of Southern Denmark

eckhard.bick@mail.dk

Background

- Spoken language data are difficult to obtain in large quantities (very time & labour consuming)
- Hypothesis: Certain written data may approximate some of the linguistic features of spoken language
  - Candidates: chat, e-mail, broadcasts, speech and discussion transcripts, film subtitle files
- This paper discusses data, tools, pitfalls and results of such an approach:
  - suitable corpora (from the CorpusEye initiative at SDU)
  - suitable tokenization and annotation methodology (CG)
  - linguistic insights and cross-corpus comparison

The corpora

- **Enron E-mail Dataset**: corporate e-mail (CALO Project)
- **Chat Corpus** 2002-2004 (Project JJ)
  - (a) Harry Potter, (b) Goth Chat, (c) X Underground, (d) Amarantus: War in New York
- **Europarl** - English section (Philipp Koehn)
  - transcribed parliamentary debates
- **BNC** (British National Corpus)
  - split in (a) written and (b) spoken sections

Grammatical Annotation

- Constraint Grammar (Karlsson et al. 1995, Bick 2000)
  - reductionist rules, tag-based information
  - rules remove, select, add or substitute information

```
REMOVE VFIN IF
    (*-1C PRP BARRIER NON-PRE-N)
    ((0 N) OR (*1C N BARRIER NON-PRE-N))
```

- EngGram (CG3 style grammar)
  - modular architecture: morphological analysis --> disambiguation --> syntactic function --> dependency
  - CG3: integrates local statistical information, uses unification
  - robust and accurate (F-pos 99%, F-syn 95% on news text)

CG adaptations for orality

- even a robust parser will suffer a performance decrease when ported from written to data with oral language traits
- CG does not need hand-corrected training corpora (which would be hard to find cross-domain, or with unified tagset)
- CG guarantees complete cross-domain compatibility, while at the same time allowing specific and repeated domain adaptations
  - Imperatives --> context rules & lexical statistics
  - Questions --> context rules
  - oral genre-specific items: interjections, emoticons (smileys)
--> lexicon additions (e.g. *grg, oy*)
--> heuristics for "productive" interjections (e.g. *oh ooh oooh, uh uh-uh*)
  - 1. and 2. pronoun frequency, "I"-disambiguation

Imperative vs. infinitive and present tense

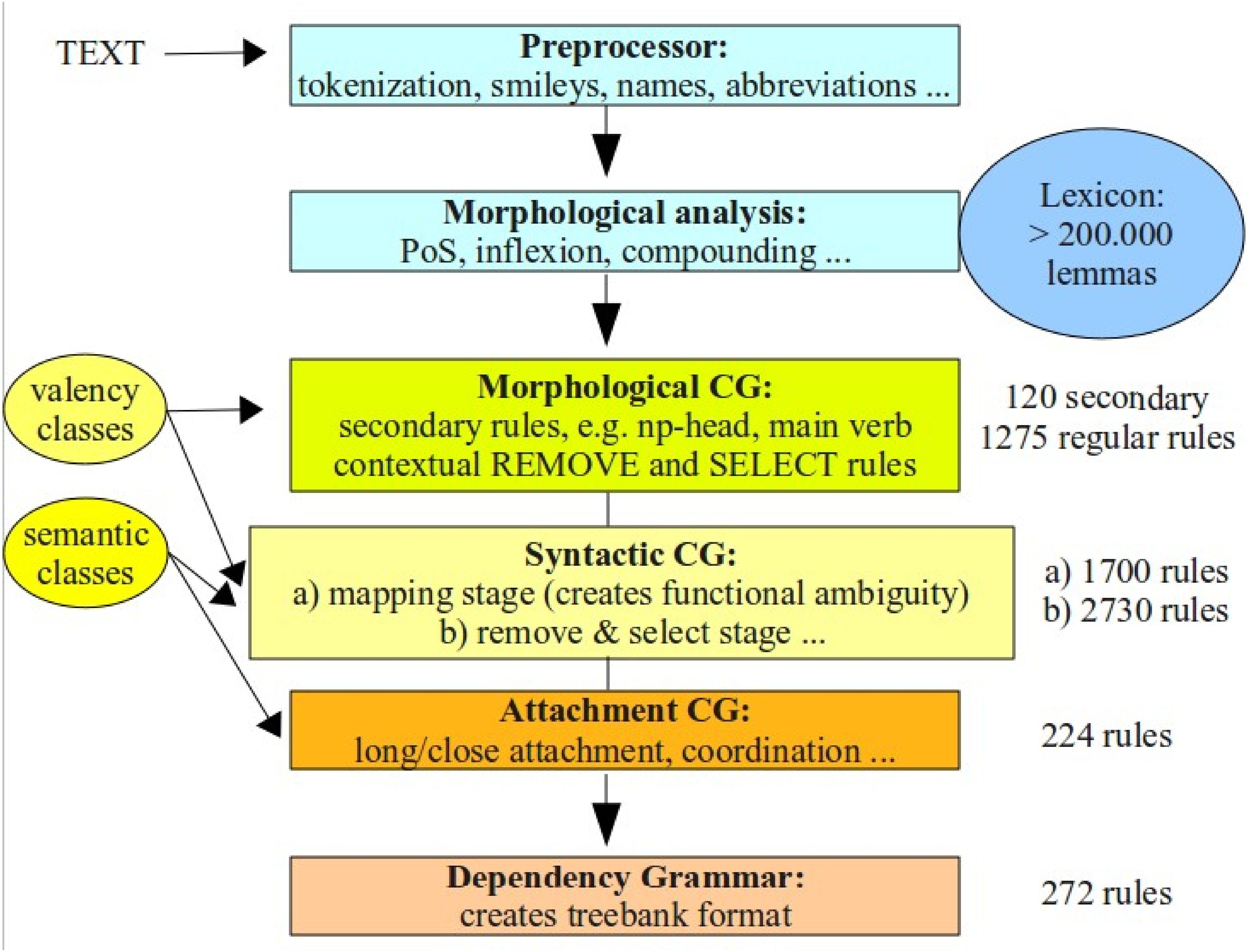
- written language parsers have an anti-imperative bias
- use context to disambiguate imperatives more precicely

```
SELECT (IMP) IF
  (-1 KOMMA) (*-2 VFIN BARRIER CLB
  LINK *-1 ("if") BARRIER CLB OR VV LINK *-1 >>> BARRIER NON-ADV/KC)
```

- use lexical likelihood statistics from mixed corpora
  - "<add>"
    - "add" <fr:12> V IMP
    - "add" <fr:68> V PR -3S
    - "add" <fr:20> V INF
  - "<achieve>"
    - "achieve" <fr:0> V IMP
    - "achieve" <fr:4> V PR -3S
    - "achieve" <fr:96> V INF

Parsing architecture

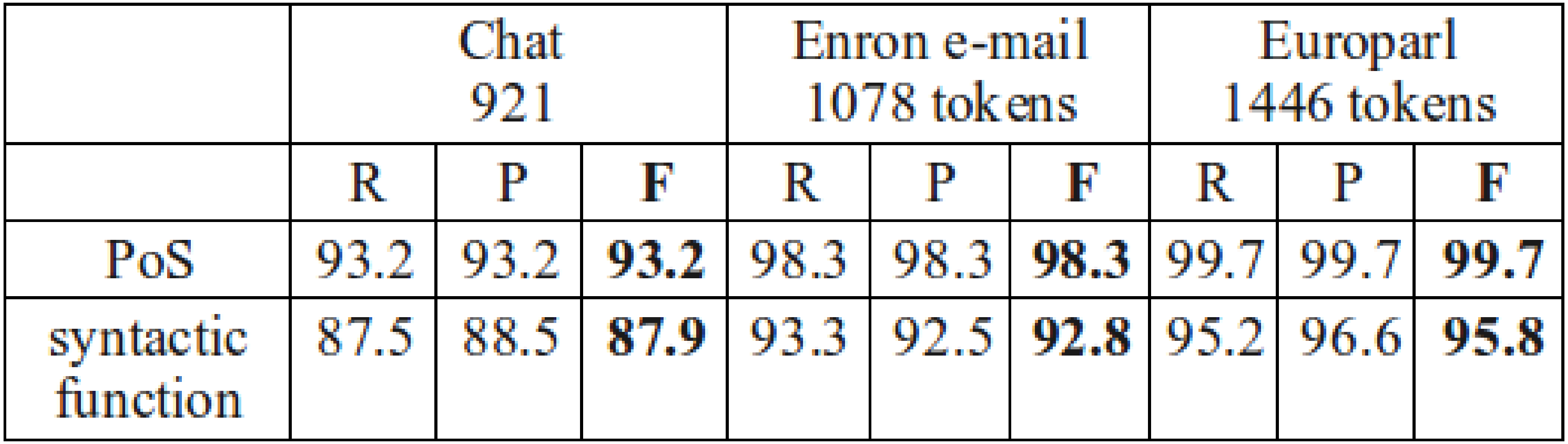
- multiple modularity
  - emoticon etc. preprocessing + morphological analysis + CG
  - multi-stage CG with rule sets at progressive levels with different annotation tasks
  - within each level: rule batches with increasing heuristicity, i.e. safe rules first: 1-2 ... 1-2-3 ... 1-2-3-4 ... 1-2-3-4-5 etc.
- lexicon support at all levels, both pos and syntax
  - valency: <vt>, <+on>, <+INF>, <vtk+ADJ>
  - semantic prototypes for nouns <Hprof>, <tool> and some adjectives <jnat> (nationhood), <jgeo> (geographical)
- highest level in this project is a kind of live dependency treebank, with all words linked to other words



TEXT
Preprocessor:
tokenization, smileys, names, abbreviations ...
Morphological analysis:
PoS, inflexion, compounding ...
Lexicon:
> 200.000
lemmas
valency
classes
semantic
classes
Morphological CG:
secondary rules, e.g. np-head, main verb
contextual REMOVE and SELECT rules
120 secondary
1275 regular rules
Syntactic CG:
a) mapping stage (creates functional ambiguity)
b) remove & select stage ...
a) 1700 rules
b) 2730 rules
Attachment CG:
long/close attachment, coordination ...
224 rules
Dependency Grammar:
creates treebank format
272 rules

Cross-corpus parser evaluation

- pilot evaluation with small data sets
- "soft" gold standard, created from parser output rather than from scratch, no multi-annotator cross-evaluation



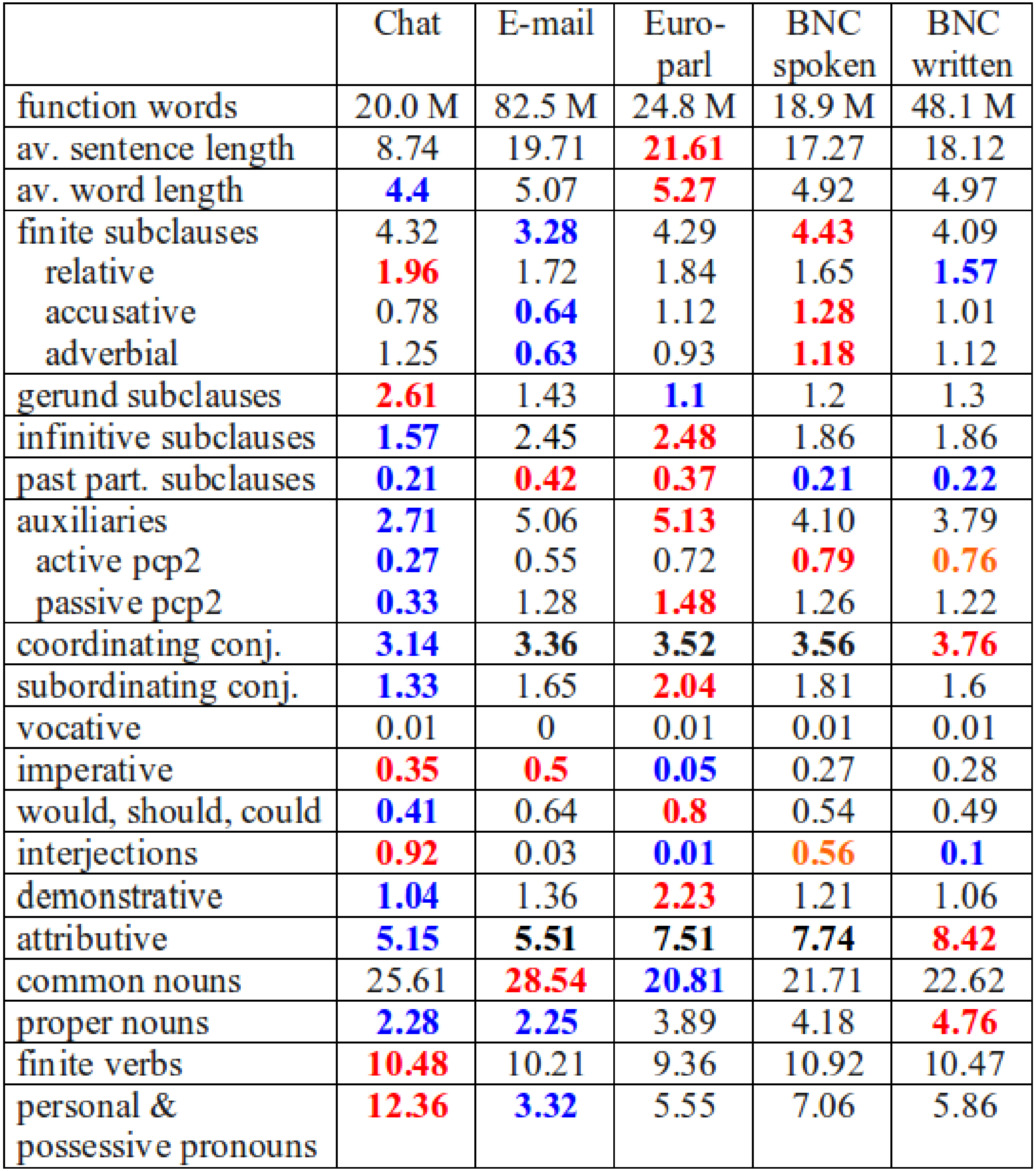
	Chat 921			Enron e-mail 1078 tokens			Europarl 1446 tokens		
	R	P	**F**	R	P	**F**	R	P	**F**
PoS	93.2	93.2	**93.2**	98.3	98.3	**98.3**	99.7	99.7	**99.7**
syntactic function	87.5	88.5	**87.9**	93.3	92.5	**92.8**	95.2	96.6	**95.8**

Problems with oral-specific traits (especially chat corpus)

- Contractions:
  - *dont, gotta*
- "phonetic" writing:
  - *Ravvvvvvvvveeee*
- unknown or drawn-out interjections read as nouns:
  - *tralalalala*
- unknown non-noun abbreviations
  - *sup (adjective), rp (infinitive), lol (interjection)*
- Subject-less sentences
  - *dances about wild and naked* ('dances' misread as noun)

Cross-corpus comparison of orality markers

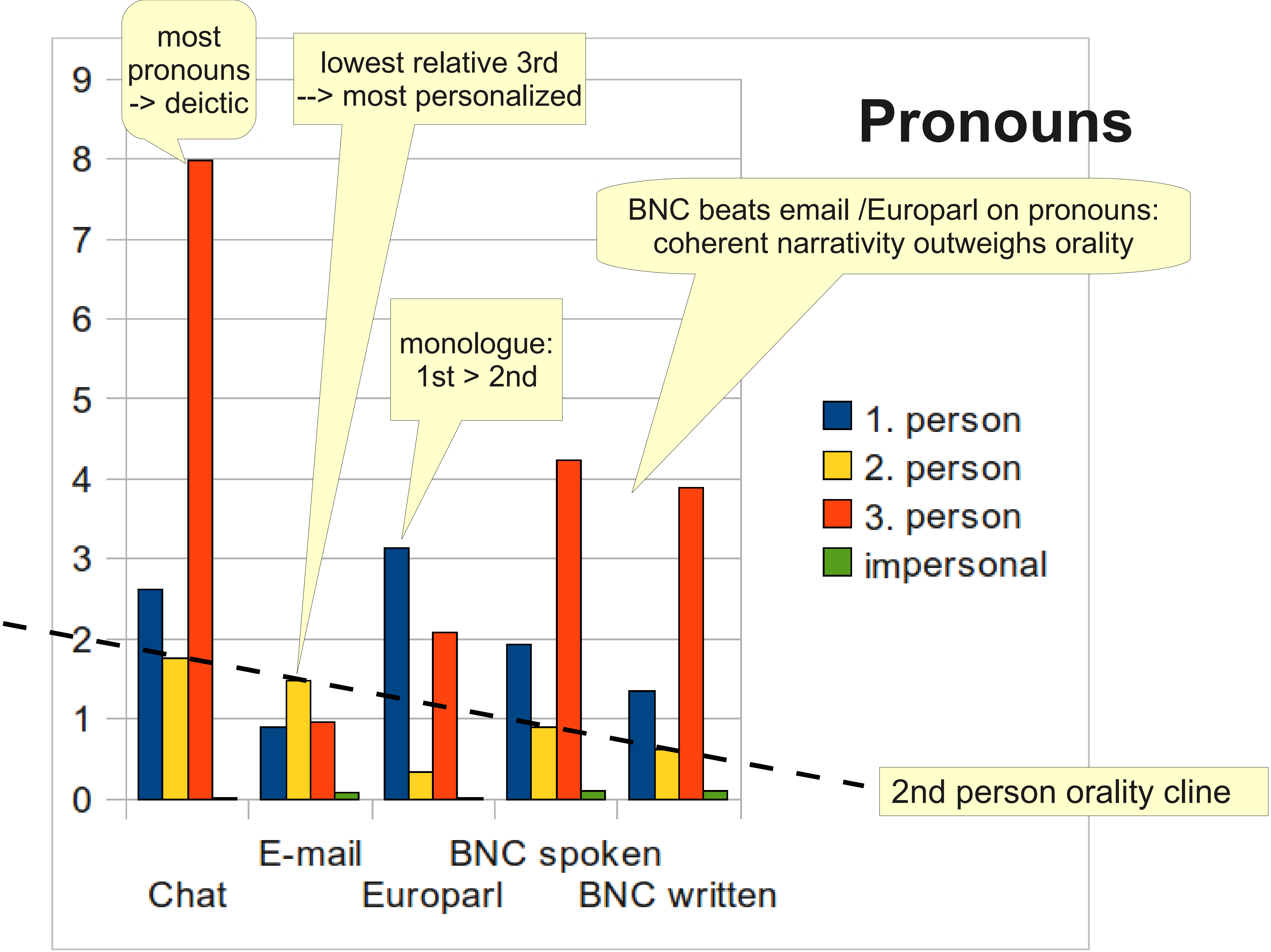
- because CG annotation is token based at all levels, even higher-level syntactic information can be used
- BNC-written included as a kind of reference corpus for the orally-influenced text types
- expected differences along a "linguistic complexity" axis:
  - **chat < e-mail < Europarl < BNC-oral < BNC-written**
- high-complexity markers:
  - verb chain length, sentence length, subordination / subclauses, would/should-distancing, passive/active ratio for participles
- low-complexity markers:
  - interjections, pronouns



	Chat	E-mail	Euro-parl	BNC spoken	BNC written
function words	20.0 M	82.5 M	24.8 M	18.9 M	48.1 M
av. sentence length	8.74	19.71	**21.61**	17.27	18.12
av. word length	**4.4**	5.07	**5.27**	4.92	4.97
finite subclauses	4.32	**3.28**	4.29	**4.43**	4.09
relative	**1.96**	1.72	1.84	1.65	**1.57**
accusative	0.78	**0.64**	1.12	**1.28**	1.01
adverbial	1.25	**0.63**	0.93	**1.18**	1.12
gerund subclauses	**2.61**	1.43	**1.1**	1.2	1.3
infinitive subclauses	**1.57**	2.45	**2.48**	1.86	1.86
past part. subclauses	**0.21**	**0.42**	**0.37**	**0.21**	**0.22**
auxiliaries	**2.71**	5.06	**5.13**	4.10	3.79
active pcp2	**0.27**	0.55	0.72	**0.79**	**0.76**
passive pcp2	**0.33**	1.28	**1.48**	1.26	1.22
coordinating conj.	**3.14**	**3.36**	**3.52**	**3.56**	**3.76**
subordinating conj.	**1.33**	1.65	**2.04**	1.81	1.6
vocative	0.01	0	0.01	0.01	0.01
imperative	**0.35**	**0.5**	**0.05**	0.27	0.28
would, should, could	**0.41**	0.64	**0.8**	0.54	0.49
interjections	**0.92**	0.03	**0.01**	**0.56**	**0.1**
demonstrative	**1.04**	1.36	**2.23**	1.21	1.06
attributive	**5.15**	**5.51**	**7.51**	**7.74**	**8.42**
common nouns	25.61	**28.54**	**20.81**	21.71	22.62
proper nouns	**2.28**	**2.25**	3.89	4.18	**4.76**
finite verbs	**10.48**	10.21	9.36	10.92	10.47
personal & possessive pronouns	**12.36**	**3.32**	5.55	7.06	5.86

- **Chat** data is most consistently oral
- **Europarl/Enron** > BNC for aux, passive pcp and would/should --> complex oral style
- **Europarl** =monologue
  - longest w and s
  - subordination
  - inf / pcp - clauses
- **BNC** oral ~ written
  - only small differ.
  - high active pcp --> narrative adj and prop --> descriptive

Pronouns



most pronouns -> deictic

lowest relative 3rd --> most personalized

monologue: 1st > 2nd

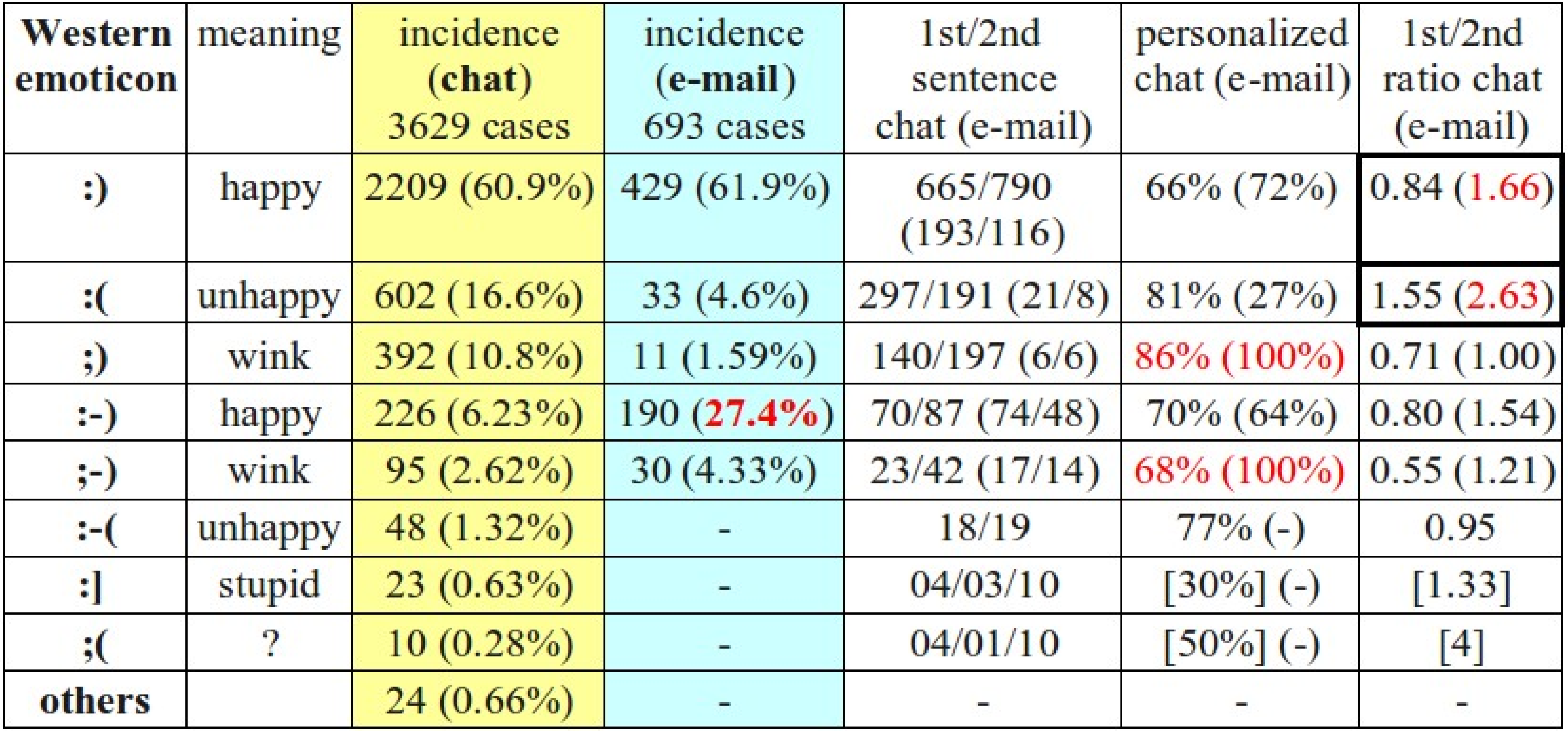
BNC beats email /Europarl on pronouns: coherent narrativity outweighs orality

2nd person orality cline

Emoticons

- high incidence, especially in the Chat corpus
- Western-tilted rather than Japanese-horizontal or number-letter-integrating
- Preprocessed as tokens rather than punctuation
- Functionally treated as free or bound adverbials
- Happy smileys are most common
  - unnosed :) more than nosed :-)
  - chat > e-mail
  - if few smileys are used, the proportion of the common ones will rise

Emoticon statistics



Western emoticon	meaning	incidence **(chat)** 3629 cases	incidence **(e-mail)** 693 cases	1st/2nd sentence chat (e-mail)	personalized chat (e-mail)	1st/2nd ratio chat (e-mail)
:)	happy	2209 (60.9%)	429 (61.9%)	665/790 (193/116)	66% (72%)	0.84 (1.66)
:(	unhappy	602 (16.6%)	33 (4.6%)	297/191 (21/8)	81% (27%)	1.55 (2.63)
;)	wink	392 (10.8%)	11 (1.59%)	140/197 (6/6)	86% (100%)	0.71 (1.00)
:-)	happy	226 (6.23%)	190 (**27.4%**)	70/87 (74/48)	70% (64%)	0.80 (1.54)
;-)	wink	95 (2.62%)	30 (4.33%)	23/42 (17/14)	68% (100%)	0.55 (1.21)
:-(	unhappy	48 (1.32%)	-	18/19	77% (-)	0.95
:]	stupid	23 (0.63%)	-	04/03/10	[30%] (-)	[1.33]
;(	?	10 (0.28%)	-	04/01/10	[50%] (-)	[4]
others		24 (0.66%)	-	-	-	-

- most personalized (1./2. person sentences) are **winks** ;) and ;-)
- unhappy smileys more speaker-marker, happy smileys more listener-marked (bold square)
  I am sad :(and you are nice :) ....not: I am nice :) and you are sad :(
- Enron more conservative than Chat: few non-happy smileys, few abbreviated smileys
- similar personalization and similar realtive distribution in 1st/2nd ratios (bold square), but fewer 2nd person smileys in e-mails

Conclusions

- We have seen that certain types of oral language features can be examined and quantified in certain types of text corpora rather than traditional transcribed speech corpora, provided that problems such as emoticons, interjections and imperatives are treated reliably
- Constraint Grammar is a robust method to handle the annotation of such corpora across varying domains
- Distribution of orality markers is neither uniform nor consistently bundled across corpus types
  - Chat data is most consistenly "oral"
  - E-mail is most personalized, but more "written" than chat - reminiscent, in fact, of traditional letters
  - Europarl as formal spoken monologue has some features that are more "written" than ordinary text
  - Some literary sources of spoken language (plays and radio in the BNC?) are not as "oral" as one would expect

Outlook

- Given the clear inter-corpus differences, a detailed error analysis should be performed, not least for the chat corpus
- Genre-specific rule modules could be added to the general grammar based on such error analysis
- Existing rules should be able to reference to a text type meta-tag for genre localization
- For the chat corpus, it would make sense to work with two orthographic levels to facilitate the use of a general parser (cp. historical corpus annotation Bick & Módolo 205)
    - (a) "as is"
    - (b) normalized [written] orthography

corpuseye

eckhard.bick@mail.dk

Parsers: http://beta.visl.sdu.dk
Corpora: http://corp.hum.sdu.dk