View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Waseda University Repository

Degrees of Orality in Speech-like Corpora: Comparative Annotation of Chat and E-mail Corpora

Eckhard Bick

Institute of Language and Communication, University of Southern Denmark
Campusvej 55, DK 5230 Odense M
eckhard.bick@mail.dk

Abstract. This paper describes and evaluates the automatic grammatical annotation of a chat and an e-mail corpus of together 117 million words, using a modular Constraint Grammar system. We discuss a number of genre-specific issues, such as emoticons and personal pronouns, and offer a linguistic comparison of the two corpora with corresponding annotations of the Europarl corpus and the spoken and written subsections of the BNC corpus, with a focus on orality markers such as linguistic complexity and word class distribution.

Keywords: chat corpus, e-mail corpus, orality, parsing, Constraint Grammar, NLP

1 Introduction

Traditional speech corpora with phonetic transcription are very labour-intensive to create, involving a huge effort in data collection, sound file management and manual transcription. Automatic transcription is today an alternative, at least for languages with a mature language technology base such as English, but the method is not error free and commercial tools will produce standard orthography, not phonetic transcription. Thus, Luz et al. (2008) report transcription speeds of 22-30 words per minute, for an ASR-assisted post editing method, with a final error rate of 3.3% - 7.83%., translating into 20 man-years of work for the one-pass one-annotator transcription of a 25 million-word corpus. A third alternative - and the position taken in this paper - is to use data where people write in a speech-like fashion, without the constraints of ordinary written production, thus in fact providing their own transcriptions. This is the case for both chat- and sms-data, and to a certain degree e-mail text. The paper describes and evaluates the annotation of two such corpora, the Enron e-mail corpus and our own Fantasy chat corpus, comparing these to the written and oral sections of the BNC and the English section of the Europarl corpus which could be described as "listener-transcribed" rather than "speaker-transcribed" and also differs from our other data in representing fairly formal, parliamentary speech.

2 The corpora

The Enron corpus is a corpus of corporate e-mails, called the *Enron Email Dataset*, and made available for research by William Cohen on his website (http://www.cs.cmu.edu/~wcohen/, http://www.cs.cmu.edu/~enron/). The data was originally made public, and posted to the web, by the (US) Federal Energy Regulatory Commission during its investigation, and later prepared by the CALO Project (http://www.ai.sri.com/project/CALO).

Our chat corpus was compiled from 4 different fantasy chat logs from Project JJ (http://www.projectjj.com), administrated and made available by Tino Didriksen. The logs were collected between August 2002 and August 2004, and cover the topics (a) Harry Potter, (b) Goth Chat, (c) X Underground and (d) Amarantus: War in New York.

The Europarl corpus used here is the English part (both original and translated) of the European Parliament Proceedings Parallel Corpus 1996-2003, prepared by Philipp Koehn. The corpus was retrieved from his website at http://www.isi.edu/ ~koehn/europarl/.

The BNC (British National Corpus, http://www.natcorp.ox.ac.uk/) was split into a written and a spoken section using section source and domain information, separating traditional written texts such as news and belletristics on the one hand from meeting recordings, lectures, television discussions, medical consultations, law reports etc. on the other hand. Of course, much of this material is - just as the Europarl transcripts - of a formal character than ordinary speech and much more standardized in terms of orthography, punctuation etc than chat or even e-mail data, and thus in many respects closer to written texts than the latte - as we will try to demonstrate in ch. 4.

3 Grammatical Annotation

All four corpora were annotated within the Constraint Grammar paradigm (Karlsson et al. 1995 and Bick 2000), using an adapted version (cf. 3.1) of the author's EngGram system (http://beta.visl.sdu.dk/constraint_grammar_languages.html). Constraint Grammar (CG) parsers are rule-based systems of a largely reductionist nature in the sense that most rules work by contextually excluding morphological, syntactic or semantic readings from a list of possible readings provided by a lexicon-based analyzer or a syntactic/semantic category mapping stage. For instance, the rule below will remove a finite verb reading (VFIN) if there is an unambiguous (C) preposition (PRP) anywhere (*) to the left (-1) with nothing but (BARRIER NON) prenominal articles, determiners and adjectives (PRE-N) in between, with a second condition that the word token in question either be a noun itself (0 N) or be followed by an unambiguous noun to the right (*1C).

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

By letting the last reading of a given type survive even in the presence of input constructions the grammar was not designed to handle, a CG system achieves a certain robustness, and all text will be analyses. Thus, the English CG we were using for our annotation task, though in principle designed for written text of the news and scholarly genre, can also produce annotations for data of varying degrees of what we will call *orality*[1] in this paper - text with a certain amount of features typical of spoken language. However, even a robust written-language parser will obviously be liable to a higher error rate when confronted with spoken-language structure and category distribution, and we therefore adapted the parser on several points.

Precise lexical and grammatical adaptability is, apart from general accuracy[2], a main argument supporting the use of Constraint Grammar rather than probabilistic taggers or parsers, for which hand-corrected training corpora would be needed separately for all the different domains involved. Furthermore, even where such gold corpora can be found, they are not likely to have been produced by the same research team with unified category sets and definitions, making comparative studies difficult. By contrast, our CG approach permits us to maintain complete compatibility across domain annotations while at the same time allowing for specific and repeated domain adaptations.

3.1 CG adaptations for orality features in written speech

One of the most important categories in this regard is the imperative, which in English is ambiguous with both the infinitive and the common present tense form (with only 3. person singular marked separately). Since imperatives are rare in ordinary written text, both statistical and rule-based parsers tend to disambiguate these form in favour of infinitives and present tense, and me had to adapt the grammar accordingly. For this, we used both context conditions

[1] The term orality is used differently in different fields. It may refer to. "oral tradition" (as opposed to written tradition) in anthropology, or to a child development stage in medicine. Here, we use the term in a literal linguistic sense, meaning "related to spoken language".

[2] Most mature CG systems such as the English ENGCG (Karlsson et al. 1995) and the Portuguese PALAVRAS system (Bick 2000) achieve part-of-speech accuracies of over 99%. Error rates for syntactic function tagging vary more, but are often under 5%.

describing the restricted left-hand context of imperatives, imperative verb sets and a statistical measure for a given verb's likelihood to occur in the infinitive. The rule below selects an imperative reading after a comma, looking for a finite verb left of the comma (*-2) with no further comma in between, then linking another left search for the word "if" and finally the left sentence boundary (>>>), allowing for nothing but adverbs (ADV) and coordinators (KC) to interfere.

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

The lexical likelihood statistics was computed from annotated mixed genre corpora, and is of course not a perfect measure for the target corpora, but it is good enough to express context restrictions in heuristic rules. In the example, the frequency tags provide percentage figures for the alternative morphological readings in what is called CG cohorts for the verb word forms "add" and "achieve", where the semantics of the latter does not support imperative use, while the former is typical of e.g. recipes.

```
"<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
```

(V = verb, PR = present tense, IMP = imperative, INF = infinitive, -3S = all person-number combinations but 1st person singular, "..." = lemma base form (here identical to the word form)

Another topic to be treated in the rule body were questions, which are much more frequent in written speech corpora than e.g. news texts. For English, word order changes in questions ask for structural-topological rules rather than statistical solutions, not only because we would need a dedicated question-gold corpus for the necessary machine learning process, but also because the learned patterns would risk compromising performance for the non-question sentences in the same corpus.

Apart from grammatical issues, in order to achieve good lexical coverage for written speech corpora, it is a necessary but not sufficient condition to provide for a larger degree of spelling variation through spell check-resembling mechanisms. A potentially bigger problem, however, are lexical items entirely specific to the oral genre, such as inventive interjections and non-word units such as emoticons. Ordinary heuristics will read the former as nouns and the latter as punctuation.

We handled interjections through lexicon additions (e.g. 'grg', 'oy'), but also needed heuristics for what one could call productive interjections, especially concerning vowel lengthening ('oh' - 'ooh' - 'oooh') and reduplication ('uh' - 'uhuh' - 'uh-uh').

Emoticons, or smileys, were captured by regular expressions in the preprocessor, to be recognized by the morphological analyzer as "adverbs" (cf. chapter 4).

The personal pronoun distribution in a written speech corpus can be assumed to differ from standard texts due to the speaker's need to refer to both himself (1st person pronouns) and the listener (2nd person pronouns), and while these effects constitute exactly the kind of descriptive research question we would like to help answer with our corpora, we also had to make a few changes to the grammar to accommodate for distributional differences in the case of "I", which is ambiguous with a Roman numeral reading more common in scholarly texts, not least after

common and proper nouns, than the pronoun reading.

3.2 The parsing architecture

Our parsing system is modular not only in the sense that preprocessing is needed for the recognition of e.g. smileys, and that CG grammar needs a morphological analysis to work on, but also as to CG itself, which is a multi-stage system separated into a morphological, a syntactic and a dependency attachment stage, each of which is again subdivided into rule batches of different heuristicity such that safer rules are run before more heuristic ones. Safe rules typically ask for unambiguous contexts, therefore rule batches need to be repeated so less heuristic rules can be retried once their context has become a little less ambiguous due to other rules. Thus, for the 6 heuristicity levels in the morphological disambiguation CG, batch order will be 1 - 2 -1 -2 - 3 - 1 - 2 - 3 - 4 -1 - 2 - 3 - 4 - 5 - 1 - 2 - 3 - 4 - 5 - 6 - 1 - 2 - 3 - 4 - 5 - 6, with 1 being the safest and 6 the most heuristic level. All in all, about 6000 rules are used. Fig. 1 illustrates the modular architecture of the system:

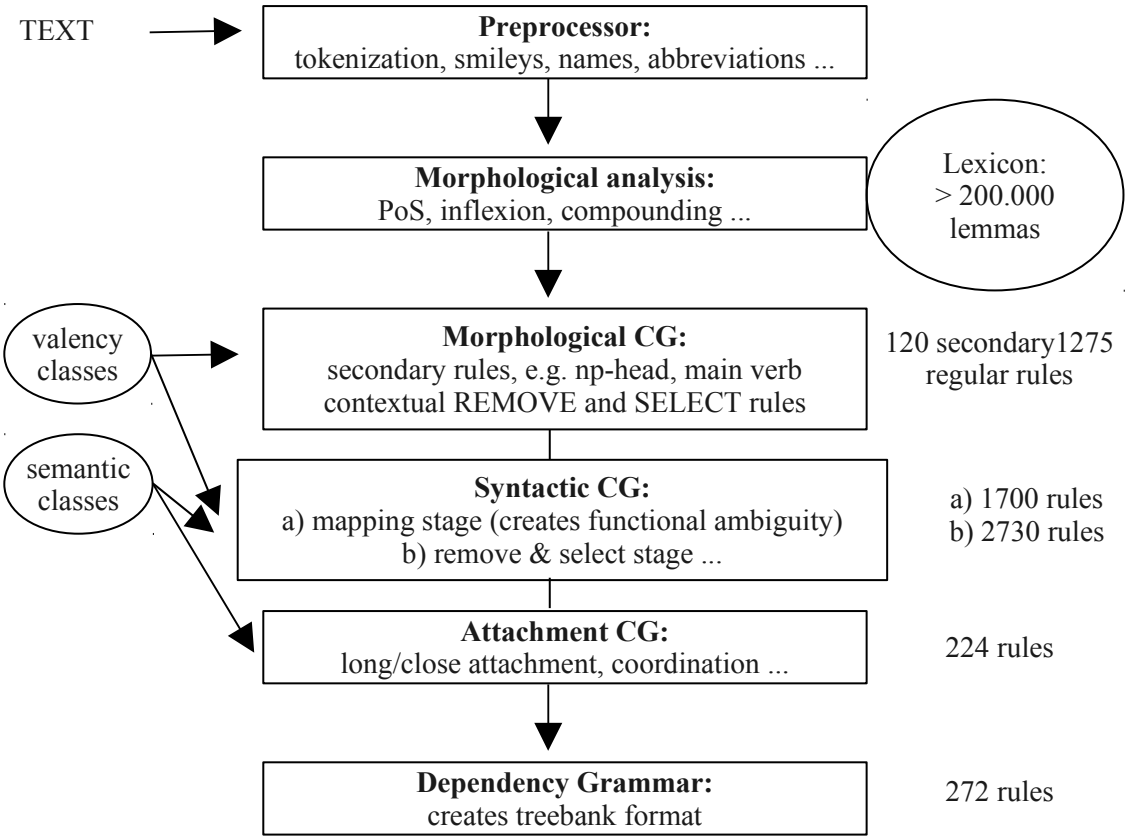

Figure 1: Parser flow chart

As can be seen from the parser flow chart, it is not only the morphological stage that is lexicon-supported. The disambiguation and structural CG modules profit from lexical information as well, in the form of so-called secondary tags, i.e. tags designed to help establish and disambiguate primary (PoS and function) tags, but not to be disambiguated themselves. These secondary tags come in two flavours:

(a) valency potentiality markers such as <vt> for transitivity, <+on> for prepositional valency, or <+INF>, <vtk+ADJ> etc for morphological selection restrictions.

(b) semantic prototypes for nouns and some adjectives, such as <Hprof> (human professional), <tool>, <jnat> (nationhood-adjective), <jgeo> (geographical adjective).

The last stage, explicit dependency links, is currently handled by a separate grammar, whose rules, however, could also be expressed within the framework of vislcg3[3], the newest compiler available for CG rules. While a corpus annotated at the dependency level can be regarded as a kind of live syntactic treebank, this extra depth was not necessary for the corpus studies presented in chapter 4, and the evaluation below will therefore focus on part of speech and syntactic function.

3.3 Cross-corpus parser evaluation

Obviously, differences in annotation accuracy can be expected - despite adaptations - when one and the same parser is used on corpora of different degrees of orality, and we therefore performed a small comparative pilot evaluation. Our method was a "soft" evaluation in the sense that gold annotations were created by manual revision of parser output rather than from scratch, and no multi-annotator cross-evaluation was used. The figures in table 1 are for function words only, considering that punctuation was not subjected to any real disambiguation, and would thus falsely "improve" results.

Table 1: Evaluation (R=recall, P=precision, F=F-score)

	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**

Concluding from these figures, the parser performed best on Europarl and worst on the chat data. Since the parser was developed for news, science and teaching texts, and is currently used and optimized for the translation of Wikipedia articles, the most likely explanation for these performance variations is the difference between the formal and professionally transcribed political jargon of Europarl on the one hand and the creative and hastily written chat texts on the other.

Error inspection did indeed reveal that the chat data in particular contained features making automatic analysis more difficult, among them orthographic and lexical unconventionalities such as

- contractions: 'dont', 'gotta'
- "phonetic writing": 'Ravvvvvvvvvvvvvveeee', 'booted'
- unknown or drawn-out interjections read as nouns: tralalalala
- unknown non-noun abbreviations: 'sup' (adjective), 'rp' (infinitive), 'lol' (interjection)

Also, subject-less sentences such as 'dances about wild and naked' led to verbs being read as nouns ('dances'), messing up the parsers syntactical analysis.

4 Comparing orality markers

4.1 General comparison

Using the annotated versions of our corpora, we have carried out a linguistically motivated

[3] Vislcg3 is an open source rule compiler (http://beta.visl.sdu.dk/cg3.html) developed and maintained by GrammarSoft ApS.

comparison of the four written speech data sets with each other on the one hand, and with the written BNC as a kind of reference text on the other. The comparison targets different levels of linguistic features, such as word class distribution, syntactic complexity and deicticity, but does so using tag-based statistics for all cases, a computationally simple and robust method made possible by the fact that CG encodes all information, even higher-level information, at the token level.

In table 2, all information with the exception of the first three rows, has to be read as corpus size-normalized percentages. High values are in bold italics, low values in bold underline[4].

Table 2: Orality markers

	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

We expected the more speech-like corpora (chat and e-mail) to be of lower linguistic complexity than the BNC reference data and the more formal Europarl, and for a number of features this is clearly the case - even though the error rates discussed above, suggest a certain margin of uncertainty.

Thus, coordination figures grow from left to right in the table, and so does the incidence of elaborating attributes (adjectives and adjectival participles). However, the chat corpus scores much more consistently along the complexity axis than the e-mail corpus and Europarl. Thus, the chat corpus has the highest occurrence of interjections[5] and pronouns, and the lowest score for verb chain length (auxiliaries), as well as for subordination, infinitive/participle subclauses

[4] Given the relatively low annotation error frequencies presented in table 1, it is reasonable to expect that valid relative comparisons between our corpora can be made without human annotation revision, even assuming a slightly unequal error distribution across the corpora.

[5] Given the fact that non-recognition of interjections was one of the problems the parser had with the chat corpus, this difference is likely to be even more pronounced than indicated.

and would/should distancing. The e-mail corpus and Europarl, on the other hand, do not consistently score in the middle between the chat corpus and the BNC. They have, for instance, more auxiliaries than the latter, and a higher passive/active ratio for participles, both of which could be interpreted as a higher level of abstraction. This is especially evident in the case of the "...ould" auxiliaries, generally implying a "reality-distance".

The Europarl corpus, in particular, is atypical for speech, and in many regards closer to running text, most likely a consequence of it consisting of formal monologue, with an abstract public in mind rather than an individual turn-taker. Thus, the Europarl data boasts the longest words and longest sentences[6], and scores highest for subordination and infinitive subclauses, as well as the rare past participle subclauses, all of which considerably complicate syntactic trees. Conversely, the chat and e-mail corpora stick together on with regard to two important orality markers, imperatives and proper nouns, scoring high on the former and low on the latter.

With the exception of interjections and the personal pronoun pattern (cf. below), there is no clear difference in orality between the two partitions of the BNC, possibly because of its high proportion of literature samples in both parts. Characteristic for the BNC as a whole is the relatively high value of active participles ('has done', 'has made') consistent with both narrative and news quotes. The BNC also scores high on "descriptivity", with high figures for attributes and proper nouns, and here, a degree difference between the spoken and written subsections can be noted.

4.2 Pronouns

Personal and possessive pronouns are conspicuously frequent in the chat corpus and relatively rare in the Enron e-mail corpus, symptomatic of the more deictic nature of the former and the fact that an e-mail lacks the narrative context of either the fantasy chat and avatar chat room, or the long coherent literary BNC texts. At least in terms of 3rd person pronouns this can also be said of the Europarl corpus, which consists of isolated monologues, and employs a more abstract and elaborative style, which is also compounded by a high attribute/noun ration, and the high incidence of (pronoun-compensating) demonstratives. The most interesting pronominal findings, however, concern not the overall pronoun figures, but their relative person distribution (fig 2, token percentages).

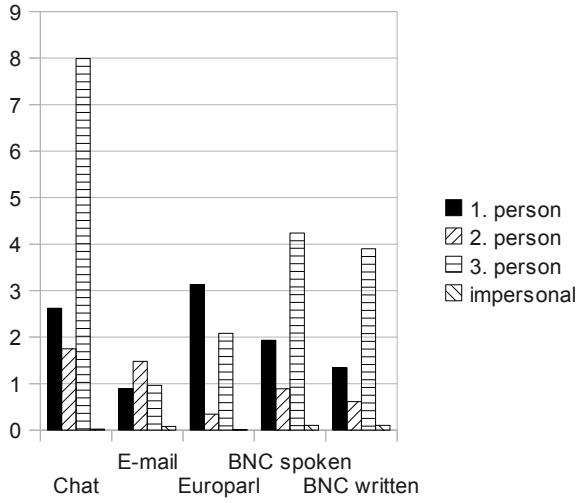

Figure 2: Person distribution of pronouns

[6] In the Europarl corpus, sentence length may also have been influenced by the fact that some of the English material is translated from Romance languages (average Europarl sentence length 32.7 words), while English itself is close to the Germanic average (24.9 words per sentence in Europarl).

It is the distribution of 2nd person pronouns that best describes our postulated cline from most oral to least oral text type. with almost a factor 3 difference between chat data and written BNC. 1st person pronouns, on the other hand, though one might expect a corresponding distribution, present two surprises. First, formal "Eurospeak" jargon has both the highest absolute use of 1st person, and the lowest absolute use of 2nd person, suggesting a monologue style addressing a non-specified (mass media) audience rather than the audience physically present. Second, the Enron e-mail corpus has the lowest absolute use of 1st person, making it the only subcorpus with more 2nd than 1st person pronouns. Seen on the background of overall pronoun use, however, the e-mail corpus 1st person usage is *not* low - in fact, due to the low incidence of 3rd person pronouns, the e-mail corpus can be said to be the *most personalized* text in pronominal terms.

4.3 Emoticons

Speakers/writers in our corpora, especially the chat corpus, made frequent use of emoticons, which would have been split and read as ordinary punctuation markers by the un-enhanced parser. Our adapted version uses regular expressions at the preprocessing stage to fuse emoticons into tokens, which are then tagged as adverbials by the parser itself. We focused on traditional Western "tilted" emoticons, not the Japanese-style horizontal emoticons or more creative letter- and number-incorporating emoticons, both of which were rare in our corpus and not covered by the tokenization process. Functionally we treated emoticons as adverbials (either free or verb-bound), the category most in line with position in the sentence, and least likely to interfere with a syntactic tree-generation module. The frequency distribution for the most popular emoticons places happy smileys at the top, accounting for about 2/3 of all cases in the chat corpus, and for almost 90% of all cases in the e-mail corpus. The short "nose-less" happy smiley :) was much more common than the "nosed" happy smiley, :-) , especially in the chat corpus:

Western emoticon	meaning	incidence (chat) 3629 cases	incidence (e-mail) 693 cases	1st/2nd sentence (chat)	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%)	-	-	-	-

The statistics also correlates emoticons with personalized sentences - defined as sentences containing 1st or 2nd person pronouns or inflexions, establishing that the most personalized emoticons are winks, which almost always constitute a direct communicative signal rather than just a statement valorization, with an average of 82.5% for short and nosed winks together in the chat corpus and 100% for e-mails. We also noted a marked difference in 1st/2nd person correlations for the chat corpus, with unhappy smileys being far more "speaker-marked" than happy smileys which are more "listener-marked", in terms of personal pronoun use. In other words, a chatter is more likely to say "I am sad :(and you are nice :)" than "I am nice :) and you are sad :(".

The Enron e-mail corpus, in spite of its larger size, contains considerably fewer emoticons, and exhibits a more conservative usage in the sense that happy smileys are far more dominant

than in the chat corpus, with a larger proportion (2:5) of the "unabbreviated" :-) smiley than in the chat corpus (1:10), and with the rarer emoticons being altogether absent. Also, while exhibiting a similar degree of personalization, and a similar relative distribution in 1st/2nd ratios (i.e. happy vs. unhappy), the e-mail writers appeared to be much more reluctant to use emoticons in 2nd person sentences than the chatters.

5 Conclusions and Outlook

We have shown how a rule-based general-purpose parser can be used and adapted to written speech corpora, annotating five corpora representing different genres, with a special focus on chat data. The grammatical annotation allowed us to demonstrate systematic differences in various orality markers across these corpora. While the chat corpus consistently scored high on a number of orality markers, both the Enron e-mail data and the Europarl parliamentary transcripts proved to be atypical as representative sources of spoken language data, with - for instance - a low pronoun count in the former and a very high degree of linguistic complexity in the latter.

Given the clear inter-corpus differences we documented, and the high chat corpus error rate in particular, it would make sense to perform a detailed error analysis of our annotation, which in turn would allow us to genre-adjust the analysis lexicon, and permit the introduction of genre-specific rules into the parser in order to improve its performance. For the chat corpus, it would also make sense to operate with two orthographic levels, one "as is" and one with normalized written orthography as suggested by Bick & Módolo (2005) for the grammatical annotation of historical data, which may suffer from spelling variations in a similar way.

References

Bick, Eckhard. 2005. Turning Constraint Grammar Data into Running Dependency Treebanks, In: Civit, Montserrat & Kübler, Sandra & Martí, Ma. Antònia (red.), *Proceedings of TLT 2005 (4th Workshop on Treebanks and Linguistic Theory, Barcelona, December 9th - 10th, 2005)*, pp.19-27

Bick, Eckhard, Marcelo Módolo. 2005. Letters and Editorials: A grammatically annotated corpus of 19th century Brazilian Portuguese. In: Claus Pusch & Johannes Kabatek & Wolfgang Raible (eds.) *Romance Corpus Linguistics II: Corpora and Historical Linguistics (Proceedings of the 2nd Freiburg Workshop on Romance Corpus stics, Sept. 2003)*. pp. 271-280. Tübingen: Gunther Narr Verlag.

Bick, Eckhard. 2000. *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* Aarhus University Press, Århus

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, and Arto Anttila, editors. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Natural Language Processing, No 4. Mouton de Gruyter, Berlin and New York

Klimt, B. and Y. Yang. 2004. Introducing the Enron Corpus. *First Conference on Email and Anti-Spam* (CEAS), Mountain View, CA. Retrieved May 29, 2010, from ftp://ftp.research.microsoft.com/users/joshuago/conference/papers-2004/168.pdf

Luz, Saturnino, Masood Masoodian, Bill Rogers and Chris Deering. 2008. Interface design strategies for computer-assisted speech transcription. Proceedings of the 20th Australasian Conference on Computer-Human Interaction, Cairns, Australia. pp. 203-210. ACM: New York.