The University of Nottingham

Applied Pragmatics: corpus-based methods and computational tools

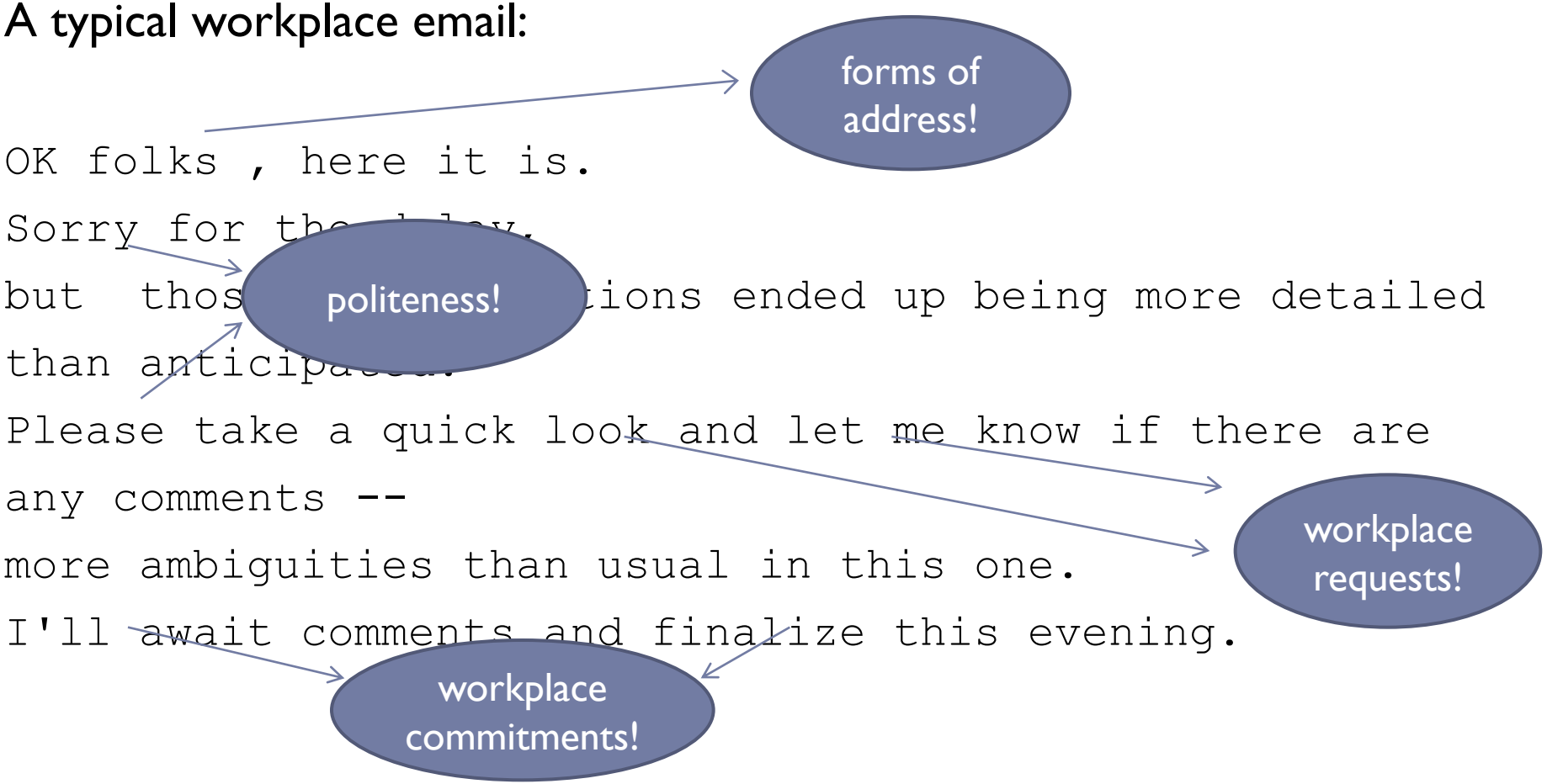
Rachele De Felice
Centre for Research in Applied Linguistics, University of Nottingham

CRAL
Centre for Research in Applied Linguistics

A typical workplace email:

```
OK folks , here it is.
Sorry for the delay,
but  those pesky questions ended up being more detailed
than anticipated.
Please take a quick look and let me know if there are
any comments --
more ambiguities than usual in this one.
I'll await comments and finalize this evening.
```

A typical workplace email:



OK folks , here it is.

Sorry for the delay,

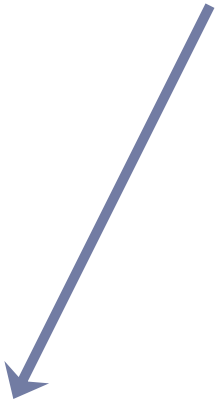
but those questions ended up being more detailed than anticipated.

Please take a quick look and let me know if there are any comments --

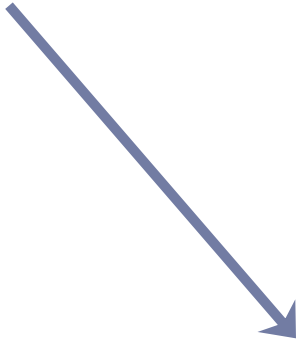
more ambiguities than usual in this one.

I'll await comments and finalize this evening.

What do speech acts look like?



to a computer



to a person

Why should we do this?

- To manage email load (e.g. task identification)
- To develop tools (e.g. corpus analysis)
- To further research on email data and pragmatics
  - Understand message content
  - Analyse style and language
- To gain insights into the language used in this domain (PROBE – the PRagmatics Of Business English)

- How do we identify speech acts?
- How do we analyse the data?
- How does this help researchers, teachers, and learners?

Related work

	Data	Classes	Features	Best results
Leuski 2005	real emails from colleagues (500)	6 (incl. 3 requests)	1-3 ngrams	P 87%, R 82% (4 classes)
Carvalho and Cohen 2006	realistic emails created for course (1716)	6	1-5 ngrams	87% acc.
Goldstein and Sabin 2006	real emails from authors' collection (280)	12	lexical items, verb classes, email specific terms, punctuation, POS tags	P 63% (1 class)
Lampert et al. 2010	emails from Enron corpus (505)	1 (requests)	wh-words, ngrams, length, modals, email specific terms	84% acc.

Table 1: Overview of related work. P = precision, R = recall

The corpus

The ideal corpus:

- Tidy – no duplicates, one sentence per line, no noise
- Wide range of topics and speakers
- Broad representation of industries and workplaces
- Current and up-to-date

The corpus

Enron emails not created as a corpus:

- Not clean
- Not filtered
- No metadata
- Automated messages and stock reports
- Private messages
- Single company
- Hard to understand

→ many of these problems arise in any email corpus

The corpus

Since both of these events happened in 1995, I would think they should not accrue in 1994 because the asset was not actually impaired in 1994, in fact the asset was fully functional and operational for the entire financial reporting period. I believe they should disclose that the possibility of expropriation exists in 1994, but I think they should not accrue.

The annotation

Research needs
Interest in detail

Data sparseness
Annotator time & effort

The annotation

- Traditional speech act categories: directives, commissives, expressives, declarations, representatives

- PROBE's 7 categories:
  - Direct request (DR)
  - Question-request (QR)
  - Open question (QQ)
  - First person commitment (FPC)
  - First person expression of feeling (FPF)
  - First person other (FPO)
  - Other statements (OT)

The annotation

- Traditional speech act categories: directives, commissives, expressives, declarations, representatives

- PROBE's 7 categories:
  - Direct request (DR) ≈ directive
  - Question-request (QR) ≈ directive
  - Open question (QQ) ≈ ??
  - First person commitment (FPC) ≈ commissive
  - First person expression of feeling (FPF) ≈ expressive
  - First person other (FPO) ≈ representative
  - Other statements (OT) ≈ representative

OK folks , here it is. → 3rd person statement

Sorry for the delay, → expression of feeling

but those pesky questions ended up being… → 3rd pers. statement

Please take a quick look and let me know if…→ direct requests

more ambiguities than usual in this one. → 3rd person statement

I'll await comments and finalize this evening. → commitment

The annotation

Statement or commitment?

- We will indicate that the utilities are hesitant to enter into these contracts.
- We continue to consider every option available to us.
- We are taking steps to be prepared to isolate the TMS system.

Embedded clauses

- I think the results are due out today.
- I know Mark is working on the report.

The annotation

"Did you remember to do everything I asked, even the small things I said in passing that didn't sound like real requests?"

The annotation

The role of *should*:

- Companies 62 and 370 should remain unchanged, thanks.
- This should include additional costs for the contract.

The role of *need*:

- I need an estimate of the cost.
- He needs to know the date of the meeting.

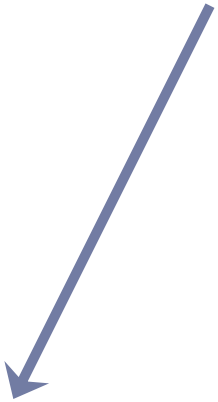
The annotation

- Currently have about 20,700 speech acts annotated
  - 263,100 words
  - ~30hrs per annotator
  - agreement ~71% - kappa 0.78

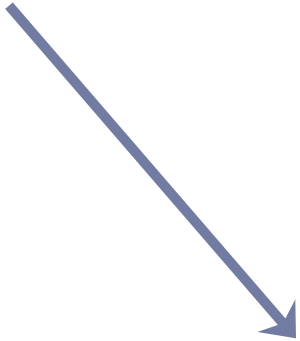
- 58% statements (1st and 3rd person)

  41.41% **OT** – other
  16.54% **FPO** – first person other
  13.62% **DR** – direct request
  10.49% **FPC** – first person commitment
  7.89% **FPF** – first person feeling
  7.31% **QQ** – open question
  2.74% **QR** – question-request

What do speech acts look like?



to a computer



to a person

The analysis

- Please pay Gaby if you see her --
- Perhaps you will want to emphasize this as being a computer system programming mistake
- There is a FERC case that you might want to also cite.
- Please let me know by Friday, if you will be able to attend
- Edit at will...
- Please provide any comments asap
- If you need immediate assistance, please contact Ava
- If you have any issues please let me know by lunchtime.
- Let me know if an account number should appear on the check.

The analysis: words and clusters

- **Please** pay Gaby if **you** see her --
- Perhaps **you** will want to emphasize this as being a computer system programming mistake
- There is a FERC case that **you** might want to also cite.
- **Please** let me know by Friday, if **you** will be able to attend
- Edit at will...
- **Please** provide any comments asap
- If **you** need immediate assistance, **please** contact Ava
- If **you** have any issues **please** let me know by lunchtime.
- Let me know if an account number should appear on the check.

The analysis: words and clusters

- **Please** pay Gaby **if you** see her --
- Perhaps **you** will want to emphasize this as being a computer system programming mistake
- There is a FERC case that **you** might want to also cite.
- **Please let me know** by Friday, **if you** will be able to attend
- Edit at will...
- **Please** provide any comments asap
- **If you** need immediate assistance, **please** contact Ava
- **If you** have any issues **please let me know** by lunchtime.
- **Let me know** if an account number should appear on the check.

Toolkit

- Tools: many levels – WordSmith, SketchEngine, POS tagging, parsing, **speech act descriptor and tagger**

- Output: corpus and sentence level descriptions
  - Corpus: word frequencies, keywords, distinctive bigrams, POS peculiarities, speech act sequences…
  - Sentence: syntax, lexicon, pronouns, punctuation…

How do we do this?

- Feature vector representations, one per instance
- Different bundles of features = different type of sp. a.

- Combination of data-driven and theory-driven
- Lexical features vs. more abstract ones

Feature set

Feature	Example values
Punctuation	;.!?, none
Subject type	noun, pronoun, none
Subject person	1^{st}, 2^{nd}, 3^{rd}
Object type	noun, pronoun, none
Object item	if pronoun, which
Has modal	yes/no
Modal is	can, will, could, should, etc.
First word	lexical item
Last word	lexical item (excl. punctuation)
Verb type	infinitive, participle, etc.
Verb tag	VBD, VB, etc.
Sentence type	declarative, question, embedded, etc.
Wh-word	who, what, when, why, etc.
Has predicative adj.	yes/no
Syntactic structures	I+modal+inf, please+imperative, etc.
Named entities	place, time, name, org., date, money
Unigrams	set of 84
Bigrams	set of 17

Toolkit

- Tools: …**speech act descriptor/tagger**
- Sentence level descriptions: syntax, lexicon, pronouns, punctuation…

I'll get a compiled answer out later today.

- Punctuation: .
- Subject: pronoun, first person
- First word: *I*
- Last word: *today*
- Object: noun (*answer*)
- Modal: yes, *`ll*
- Complex structure: I + modal + infinitive
- Named entity: date (*today*)

Tools

- Features are obtained using C&C parser
  - POS tagger, CCG dep. parser, named entity recogniser
  - Verb and sentence type information
- Bigram list compiled from L2 and some Enron data
- Unigrams – tf-idf
- Classifiers: maxent – av. accuracy 75%
  - previous work – L2 data – SVM and maxent, both 79%

Results

	Recall	Precision	F-score
DR	75.2%	86.8%	80.6%
FPC	61.8%	68.1%	64.8%
FPF	43.3%	70.7%	53.7%
FPO	74.6%	62.1%	67.8%
OT	92.9%	81.7%	87.0%
QQ	78. 2%	78.2%	78.2%
QR	50.0%	73.7%	59.6%

Results

	Recall	Precision	F-score
DR	75.2%	86.8%	80.6%
FPC	61.8%	68.1%	64.8%
FPF	43.3%	70.7%	53.7%
FPO	74.6%	62.1%	67.8%
OT	92.9%	81.7%	87.0%
QQ	78. 2%	78.2%	78.2%
QR	50.0%	73.7%	59.6%

Results

Most frequent confusions:

- Commitments tagged as statements
- Expressives tagged as statements
- Simple questions and request-questions

Tricky sentences

- **Unrecognised requests:**

  John - please drop Mike a line about the Contracts system implementation on TW.

  Also need to begin to detail the volumes from Long term firm contract.

  Given FERC's findings, you must order refunds.

  You should probably send an invite anyway.

Tricky sentences

- **Unrecognised requests:**

  John - please drop Mike a line about the Contracts system implementation on TW.

  Also need to begin to detail the volumes from Long term firm contract.

  Given FERC's findings, you must order refunds.

  You should probably send an invite anyway.

- **Unrecognised commitments:**

  I am getting home early.

  During these dates, I could also have your phone forwarded to Ava if you wish.

  We are having a meeting in 49C2 at 3:00 regarding the TW/NNG deal.

  I'm hoping to have a conversation with Mary K. tomorrow before its posted [sic].

Tricky sentences

- **Unrecognised requests:**

  John - please drop Mike a line about the Contracts system implementation on TW.

  Also need to begin to detail the volumes from Long term firm contract.

  Given FERC's findings, you must order refunds.

  You should probably send an invite anyway.

- **Unrecognised commitments:**

  I am getting home early.

  During these dates, I could also have your phone forwarded to Ava if you wish.

  We are having a meeting in 49C2 at 3:00 regarding the TW/NNG deal.

  I'm hoping to have a conversation with Mary K. tomorrow before its posted [sic].

- **Unrecognised feelings:**

  I'm discouraged to still be sick (asthma or bronchitis or the like).

  I was disappointed that I could not attend.

  I get a bit frustrated when Stan says that we all need to re-focused back to work.

Tricky sentences

- **Unrecognised simple questions:**
  Could the gift recipient select the menu items?
  Should we attach the first round of questions?
  Can these really all be receipt imbalances?

- **Unrecognised indirect requests:**
  Will this time suit you?
  Will you be able to print them?
  Does Fri at 10 work for you?
  Do you have a few minutes to meet this morning?
  Do you happen to have an email address for Elza?
  What time is good for you?
  How about 2:30?

Top ten features (previous work)

- First Word
- Punctuation
- Sentence Type
- Syntactic Structures
- Last Word
- Wh Word
- Modal Is
- Subject Is
- Verb Tag
- Has Unigram please

…or just bag-of-words?

What can we learn?

Data representation → improved language description

- Features of workplace English:
  - Adjective use (very small set – able, available, free, interested)
  - Typical main verbs/verb + object combinations
  - Role of proper nouns

- Comparison with learner data:
  - Role of adverbs (urgently, immediately)
  - Discourse sequences

- Comparison with spoken data
  - Some technical issues

Case study: bossy adverbs

1. Note low frequency of adverbs in the data

2. Identify potential NNS dataset for comparison
   - Business emails from the Cambridge Learner Corpus (test data; © Cambridge University Press)

3. Observe key difference:

Clare, I request you to call Mr. John **immediately.**

Would you please **urgently** try to solve the problem?

Case study: bossy adverbs

Observe key difference:

Clare, I request you to call Mr. John **immediately.**
Would you please **urgently** try to solve the problem?

- Only used (rarely) by NS in commitments
  - Well we need to get on that **urgently** and find out.
- Grammatically correct, pragmatically inappropriate use of adverbs → potential topic requiring closer attention

Some linguistic thoughts so far

- Much more direct
- Incomplete mastery of the syntactic structures
  (at least at this proficiency level)
- Knowing the language is not enough – need to know how to use it, too

Many unanswered questions

- How do NS colleagues react to NNS speech acts?
- What is the wider context of the request?

  (e.g. supportive/grounding moves)
- Do they talk about the same things?
- Many more levels of analysis…

Ongoing challenges

Technical issues:

- complex language
- speech-like

→ obstacles to using some tools

```
We will tap into the contacts we have as a
company to aid terminated employees in their
job search.

Talk to you later?
```

Ongoing challenges

- How much data do we need?
  - Currently at 20,700 speech acts – 263,100 words
  - ~50% is statements
  - Time consuming to process and annotate
  - Extremely difficult to obtain consent to use email data

Ongoing challenges

- Are we asking the right questions?

  "There can be tensions between speech act classifications and taxonomies which were developed on the basis of invented examples, and the analysis of speech acts in corpus data."

  (O'Keeffe, Clancy, and Adolphs 2011)

Ongoing challenges

- Are we asking the right questions?

  "There can be tensions between speech act classifications and taxonomies which were developed on the basis of invented examples, and the analysis of speech acts in corpus data."

  (O'Keeffe, Clancy, and Adolphs 2011)

- How to code for form vs. function?
  - Two tags? Special flag?
- How much does discourse structure influence function?
  - *Jeff, if the lawyers can't,* ***I'm sure we can ask Mark to get the filing****. Let me know.*

To conclude

A speech-act tagged email corpus: new solutions, new problems?

- Two NLP tools in development
- Learning more about speech acts
- Improving linguistic description

- Much human groundwork needed
- Taxonomy is slippery
- You can never have enough data

Rachele De Felice would like to gratefully acknowledge the support received by the Leverhulme Trust.

We thank Jeannique Darby, Tony Fisher, and David Peplow for their invaluable work in manually annotating the Enron email data and identifying taxonomical issues.