A classification scheme for annotating speech acts in a business email corpus

Rachele De Felice, University College London
Jeannique Darby, University of Oxford
Anthony Fisher, University of Nottingham
David Peplow, Sheffield Hallam University

Abstract

This paper reports on the process of manual annotation of speech acts in a corpus of business emails, in the context of the PROBE project (PRagmatics of Business English). The project aims to bring together corpus, computational, and theoretical linguistics by drawing on the insights made available by the annotated corpus. The corpus data sheds light on the linguistic and discourse structures of speech act use in business email communication. This enhanced linguistic description can be compared to theoretical linguistic representations of speech act categories to assess how well traditional distinctions relate to real-world, naturally occurring data. From a computational perspective, the annotated data is required for the development of an automated speech act tagging tool. Central to this research is the creation of a high quality, manually annotated speech act corpus, using an easily interpretable classification scheme. We discuss the scheme chosen for the project and the training guidelines given to the annotators, and describe the main challenges identified by the annotators.

1 Introduction

This paper describes the development of the first version (v. 1.0) of a classification scheme for the manual annotation of speech acts in a corpus of business emails, in the context of the PROBE project (PRagmatics of Business English). The overall aim of the project is to bring together corpus, computational, and theoretical linguistics by drawing on the insights made available by the annotated corpus to gain a better understanding of the linguistic and discourse structures of speech act use in business email communication. The results of the

project enable an enhanced linguistic description of speech acts which can be compared to the theoretical linguistic representation of speech act categories to assess how well traditional distinctions relate to real-world, naturally occurring data. From a computational perspective, the annotated data is required for the development of an automated speech act tagging tool, building on previous research (De Felice and Deane 2012) based on less complex non-native English language.

Central to this research is a high quality, manually annotated speech act corpus with an easily interpretable classification scheme, the creation of which is discussed in this article. After defining the research goals of the project (Section 2), we discuss the data and classification scheme chosen for the project and the training guidelines given to the annotators (Sections 3 and 4). We describe the main challenges identified by the annotators (three of the authors of this paper), and possible solutions to these challenges (Sections 5 and 6). In Section 7, we briefly outline the use of the tagged data as a resource for linguistics and natural language processing (NLP) research.

This paper's contributions are as follows: a) the introduction of an annotated corpus resource and a speech act annotation scheme which can be adapted by the research community; b) a discussion of annotation problems and ambiguities, and related solutions adopted; c) a discussion of some methodological issues around the task of pragmatic annotation; d) the introduction of two tools using pragmatically annotated data. By highlighting both our progress in this task and the unresolved methodological questions that remain, we advocate ongoing discussion of these issues with the goal of further advancing the development and analysis of pragmatically annotated corpus resources.

2 The PROBE project: Research goals

The PROBE project uses corpus and computational linguistics to create a description of the pragmatics of Business English, in particular email communication. There are three main goals to the project. From a linguistic point of view, an analysis of the corpus data yields insights into the linguistic and discourse structures of business communication, which can then be compared to theoretical linguistic descriptions of speech acts. From a computational perspective, the annotated data is used to develop an automated speech act tagging tool (De Felice and Deane 2012) for email texts, which can contribute to the already existing tools for the automated analysis and tagging of language. Finally, from a language learning angle, this kind of linguistic knowledge can enhance the teaching and learning of Business English by providing learners with real-world models of language use (see for example De Felice 2011a).

The project responds to the call raised by McEnery and Wilson (1996), who write that "quantitative accounts…would be an important contribution to our understanding of pragmatics" (McEnery and Wilson 1996: 99). As discussed in Section 3.2, pragmatic studies often focus only on a small amount of data, analysed qualitatively, such that it is not always possible to establish the more general applicability of their findings. Our corpus annotation efforts, combined with the use of corpus analytic and NLP techniques, enable us to present the quantitative account advocated above, contributing to the growing discipline of corpus pragmatics. Although the focus of our study is the pragmatics of speech act use in workplace emails, the methodological framework we present is intentionally domain-neutral and does not use categories specific to workplace communication, so as to be applicable to other areas, too. The long-term outlook of the development of the classification scheme is to obtain a resource that can enable straightforward comparisons between different corpora annotated with the same scheme, as well as between theoretical accounts and actual instantiations of different types of speech acts. Therefore, the research questions addressed in this article are:

1. Which categories should be included in a classification scheme of speech acts for corpus annotation?
2. What is an appropriate unit of annotation for speech act tagging?
3. How can the annotation problems encountered in speech act tagging be resolved?
4. Can the speech act tagging task be automated?

3 The corpus

In this section, we discuss a number of issues related to corpus choice for our task. We first introduce the existing email corpora available, before describing the pre-processing required by the data.

3.1 Email corpora

Workplace email communication (often termed 'Business English') is an area of investigation which is becoming increasingly important as English emerges as the de facto language for international business (Gimenez 2006; Bjorge 2007; Jensen 2009; Ho 2010; Newton and Kusmierczyk 2011). Research on email has also been of interest to the NLP community for some years (for example Carvalho and Cohen 2005, 2006; Goldstein and Sabin 2006; Lampert *et al.* 2008, 2010), as will be discussed in more detail in Section 4.1, particularly in

relation to the aim of assisting email management through speech act detection, automatic thread summarization, or automated sorting. However, in both these fields the research has suffered from a lack of large corpora consisting of freely available real-world data. It is, understandably, very difficult to obtain large email collections from acquaintances or companies because of privacy and intellectual property concerns, so researchers tend to make use of datasets not available to the wider community. This means that it is impossible for the wider research community to replicate or extend their findings, or reuse the resources created.

There are currently three publicly available email datasets: the W3C Corpus, which contains mailing list emails from the W3C website (the World Wide Web Consortium, which is concerned mainly with the development of web standards); the CSIRO Corpus, which contains mailing list emails from the Australian national science agency (both used for the TREC Enterprise Track for research on automatic summarization, threading, and question answering); and the Enron email Corpus (see for example Klimt and Yang 2004; Berry, Browne, and Signer 2007). Full details about these datasets can be found in Ulrich *et al*. (2008). The first two datasets are not appropriate for research on workplace communication because their content is very technical in nature, consisting mainly of specific requests, or informal discussions of particular technical issues. Furthermore, there is little metadata available for these corpora because the tasks for which the data was collected did not require knowledge of contextual information such as the mutual relationship of the participants.

The Enron corpus, on the other hand, does not originate from a research need; it was made publicly available following the legal proceedings against the corporation. In its original form, it consists of the unedited, unmodified dump of all the employees' mailboxes, with their original folder structure unmodified. Its size (several messages words) and authenticity, together with the fact that it contains data from several different speakers, makes it an invaluable resource for the research community in several disciplines. In its original format (as distributed for example by the Linguistic Data Consortium or by William Cohen at http://www.cs.cmu.edu/~enron/) it is not immediately usable for corpus linguistics research, because of its complex nested folder structure and the presence of a large amount of extraneous material typical of emails such as headers, HTML markup, and other symbols.

However, corpus linguists can make use of the EnronSent subset of the data (Styler 2011), a section of the original corpus pre-processed specifically for the purpose of linguistic research. The data is in plaintext format, and has been cleaned of all noise such as headers, spam messages, HTML, automated mes-

sages (such as reservation confirmations), legal disclaimers, and repeated quoted and forwarded messages that result from the practice of reproducing the text of previous emails in the body of the current email, leading to a high volume of duplicate text. Styler's corpus, as the name suggests, avoids some of these problems by taking only emails from the 'sent' folders of the corpus, on the assumption that these are more likely to contain actual content of interest. He further applies a 'scrubbing' program to the data to delete as much as possible of the noise described above (cf. Styler 2011). The resulting corpus is a concatenation of plain text files, consisting of 96,100 messages (13.8 million words). There is generally no clear indication of the identity of senders and recipients; as we discuss later on, lack of knowledge about the participants and their relationships to each other can lead to serious problems during the annotation task.

3.2 Data pre-processing and segmentation

Further pre-processing was necessary for the data before the annotation task. Through a combination of ad-hoc UNIX sed scripts and manual checking we removed more extraneous material, including headers, timestamps, press releases, and so on. A further key issue in the pre-processing stage is deciding how to delimit the unit of annotation and, by extension, the unit of study. In other words, what falls within the scope of an 'utterance'? Archer *et al*. (2008) observe that "the utterance is regarded as a key unit of analysis in pragmatics, but it evades easy definition" (Archer *et al*. 2008: 633), and one of the first challenges faced in the course of this project has been identifying both a suitable definition, and a viable segmentation strategy. As this is written text, individual sentences are generally clearly delimited by punctuation and spacing. However, it is common for more than one speech act to be contained in the same sentence. We are concerned here with the formal aspect of speech acts only, that is, where the sentence contains more than one clause, not with ambiguous utterances which could be interpreted in more than one way depending on context (this will be addressed in Section 6). An example of this kind of sentence is *He'll write the report tonight and I'll forward it to you*, where we have two distinct clauses communicating two distinct speech acts – a statement and a commitment. In cases such as these, it would be appropriate to segment the two clauses into two distinct utterances, and annotate each one separately. This issue can also occur when the two utterances share a subject, as in *I'm the manager here and want you to reply straight away.* Again, here we have both a first person statement and a request, and it would be misleading to assign just one tag to the utterance, as well as difficult to decide which one to assign.

To address this issue, we need a segmentation procedure which should be transparent, automated, and repeatable, and obviously free from human bias. Most of the work discussed in the literature refers to the segmentation of spoken language, either audio data or its transcription, which is not applicable to our task (Geertzen *et al*. 2007, Lendvai and Geertzen 2007). Given the mostly well-formed nature of our text data, we chose to explore the viability of a simple automated method which relies on a Combinatory Categorial Grammar (CCG) supertagger[1] (Clark and Curran 2004; Hockenmaier and Steedman 2007). The supertagger assigns different tags to conjunctions depending on whether they are sentential or phrasal coordinates. For example, given the two sentences *I am happy but lonely*, and *I am happy but he is lonely*, the two instances of *but* are tagged differently, reflecting the different nature of their conjuncts, as shown below (other tags omitted for clarity).

(1a) I am happy but|but|CC|I-ADJP|O|conj lonely.

(1b) I am happy but|but|CC|O|O|conj he is lonely.

We ran the data through the supertagger, and post-processed the output (using a sed script) by inserting a new line character where the second type of conjunction occurred. We refer to the segmented units thus obtained as utterances. Our current script is occasionally too broad, and suffers from some overgeneralisation whereby clauses are segmented more often than required. Although it has proved suitable for our annotation so far, we are analysing the parser output further to delimit more clearly the instances where segmentation should occur. Stiles (1992) proposes that the following types of clauses should be treated as independent utterances: simple sentences; independent clauses; non-restrictive independent clauses; elements of a compound predicate; terms of acknowledgement, evaluation, or address. We plan to incorporate these indications in our revision of the segmentation script.

After the pre-processing of the data outlined above, the first phase of the PROBE project has resulted in the annotation of 263,100 words/approximately 20,700 speech acts.

4 The classification scheme: previous and current approaches

In this section, we briefly review related work in speech act annotation; we refer the reader to Archer *et al*. (2008) for a detailed discussion of the wider challenges in pragmatic annotation, including, but not limited to, speech act identification. In particular, we discuss the types of speech act classification schemes

found in the literature, and the reasons for establishing our own scheme. As mentioned in Section 2, the underlying motivation for annotating the data is to develop a resource that can allow the description of the main categories of speech act identified in the linguistics literature. By then examining the different categories, we will, firstly, acquire a better understanding of the pragmatics of business communication. Furthermore, the annotated corpus will enable comparisons with the traditional descriptions based on invented examples rather than the analysis of actual instances of language, initially in the business domain, but with the long-term goal of extending the comparison to other language domains also. Therefore, the scheme used must be broadly comparable to the traditional speech act categories of linguistics literature, and not include categories which would be too specific to a particular domain.

4.1 Speech act annotation in linguistics and NLP

An issue that frequently arises in speech act research is the distinction between locutionary and illocutionary meaning (Austin 1962). Briefly stated, the former refers to the form of the speech act, what is literally being said or written (for example, *I'm too short to reach the shelf*), while the latter refers to the function of the speech act, what the speaker is actually intending to communicate with that utterance (for example, asking for help in reaching an item on the tall shelf). As we will see in the course of this paper, how to account for this distinction is one of the main challenges of speech act annotation.

One of the most interesting speech act taxonomies to do so is the Verbal Response Mode (VRM) framework (Stiles 1992), originally designed to annotate dialogues from patient-psychologist interactions. The VRM scheme explicitly encodes the locutionary and illocutionary aspects of a speech act and the fact that the two might not coincide. It uses a set of one-letter codes for basic categories, which can be combined in different ways depending on the form-function characteristics of the utterance. In work conducted prior to this project (De Felice and Deane 2012), the categories used were modelled very closely on the VRM, but further analysis showed that it was not a good fit for our research aims and for the annotation task, in particular because it originates in the context of annotating dialogue rather than written data.

In fact, within the body of work on pragmatic annotation, there is very little work concerned with data other than spoken language. For example, the field of dialogue act classification and intention understanding for dialogue modelling, though very rich and long-established (for example Core and Allen 1997; Stolcke *et al*. 2000; Georgila *et al.* 2009), poses rather different questions from those of our research. Dialogue systems deal mainly with synchronous, not

asynchronous, communication, and the exchanges are less likely to contain complete sentences. These systems are usually designed with a limited information-seeking domain in mind, such as making travel reservations, and the categories used reflect this kind of spoken interaction. Typical categories include dialogue acts which reflect the collaborative and interactive nature of spoken language, such as backchannelling, answering immediate questions, repetitions, or answer elaboration. Although email communication is also interactive and collaborative, the exchanges are usually not immediate, and these kinds of discourse moves tend to feature much less often than complete sentences. Often these classification schemes include a very large number of categories: Stolcke *et al*. (2000), for example, include 42 basic tags. We find that having a smaller set of categories better fulfils our annotation needs, as we discuss in more detail below. Therefore, although dialogue act research does share some concerns with our work – in particular with regard to issues of category choice and access to all levels of contextual information – we cannot adopt its speech act taxonomies for our annotation task.

In recent years, there has been some work in the NLP community on automated speech act identification in emails (Carvalho and Cohen 2005, 2006; Leuski 2005; Goldstein and Sabin 2006; Lampert *et al*. 2008, 2010). However, within this body of research there is no shared annotation framework, as each research project pursues different objectives and uses different categories, depending on the topic of interest. Often these annotation schemes are too project-specific to be used for other research, as in the case of Leuski (2005) who includes several categories of requests, or too general, for example focusing only on request identification and having no categories for other kinds of speech acts (Lampert *et al*. 2010). Furthermore, much of this work assigns the tags to entire email messages rather than individual utterances.

Other projects, developed in the context of linguistics rather than NLP research, also tend to be too domain-specific for our needs. Archer (2005), for example, describes data from the judicial domain, using relevant categories such as counsel, sentence, and require. Maynard and Leicher (2006) discuss the annotation of MICASE, the Michigan Corpus of Academic Spoken English, for pedagogical purposes. Their categories are therefore focused on events that can occur in a classroom, with pragmatic features including homework assignment, explaining terms, evaluations, and tangents. Finally, Kallen and Kirk (2012) do provide more general-purpose annotation, but based, again, on spoken language rather than email or written data.

As none of these schemes seem to fully respond to our requirements of being sufficiently general, non-domain specific, and easily related to traditional

speech act categories, we decided to develop our own speech act classification scheme, as discussed in Section 4.2.

4.2 Our classification scheme

Our aim is to design a viable classification scheme of speech act categories that is relevant to our project needs, easy to implement, while clearly documented and replicable. To achieve this goal, it is necessary to balance a number of competing demands. From the researcher's point of view, it would be desirable to have a richly annotated corpus, identifying a wide range of actions such as apologising, refusing, committing, complaining, inviting, requesting, and many others which are often discussed in the literature. On the other hand, a very detailed classification scheme can lead to data sparseness, whereby the examples available are not sufficient for meaningful generalisations. As we have found in our project, even with a broad categorisation there are significant disparities in category frequency, with some categories making up less than ten percent of the dataset (see Section 5.1). Furthermore, a very complex classification scheme may prove problematic for the annotators. It would increase both time needed for their training and for the annotation task (and therefore project costs), while also increasing the potential for errors and confusion.

We therefore decided on a classification scheme consisting only of seven broad categories:

a) Direct request
b) Question-request
c) Open question
d) First person commitment
e) First person expression of feeling
f) First person other
g) Other statements (second and third person)

These categories are closely aligned to those of traditional speech act theory described by Austin and Searle (directives, commissives, expressives, and representatives). The seven categories are summarised in Table 1; full guidelines are given in Appendix A, and will also be made available by the first author on her website.

Table 1: Speech act categories used

Speech act	**Tag**	**Example**
Direct request	DR	*Please send me the files.*
Question-request	QR	*Could you send me the files?*
Open question	QQ	*What time is the meeting?*
First person commitment	FPC	*I will attend the meeting.*
First person expression of feeling	FPF	*I am uncertain about the agenda.*
First person other	FPO	*I am an employee of this company.*
Other statements	OT	*The meeting is at 8 tomorrow.* *You always work so hard.*

Despite their grounding in traditional speech act theory, the PROBE categories do not have a one-to-one correspondence with the traditional categories, as we can see in the table. Directives are represented by two types of requests, 'direct request' (DR) and 'question-request' (QR). This reflects the widespread use in the literature of distinguishing between direct and indirect requests; by maintaining this distinction, it is easier to identify and analyse the different types of requests within the corpus. First person commitments (FPC) and 'expressions of feeling' (FPF) are comparable to commissives and expressives, respectively. As Table 1 shows, the former includes statements such as *I will attend the meeting*, while the latter includes any articulation of feelings of personal sentiment such as apologies, joy, congratulations, and so on.

The single category of representatives has also been divided into two categories in our scheme: 'first person other' (FPO, such as *I am an employee of this company*) and 'other statements' (OT, such as *He is the vice president* or *The meeting is at 8)*. Both categories share the function of stating facts and informing. We introduce the distinction between first person subjects and other subjects principally so that it is easier to extract and compare the different types of first person statements in the corpus. In making this distinction, we hope to gain a richer understanding of the different ways in which individuals talk about themselves in workplace email.

Finally, unlike the traditional categories, our classification scheme does not include declaratives (e.g. excommunications, christenings, etc.) as these are highly institutionalised acts unlikely to be used in everyday workplace emails. We also introduce the category of open questions (QQ, those requiring simple

answers – *what's the time*, *what's your name*, and so on). These are typically included in the directives category in discussions of speech acts in the literature, but their information-seeking nature is a distinctive and integral element of workplace discourse which warrants their specific identification to help us understand their use in this domain.

These correspondences make it easier to assess how our findings relate to theoretical linguistics claims, as outlined above. They are also sufficiently generic and non-domain specific to allow easy reimplementation within other research projects, as they do not describe actions and events that are only specific to workplace contexts.

5 The annotation process

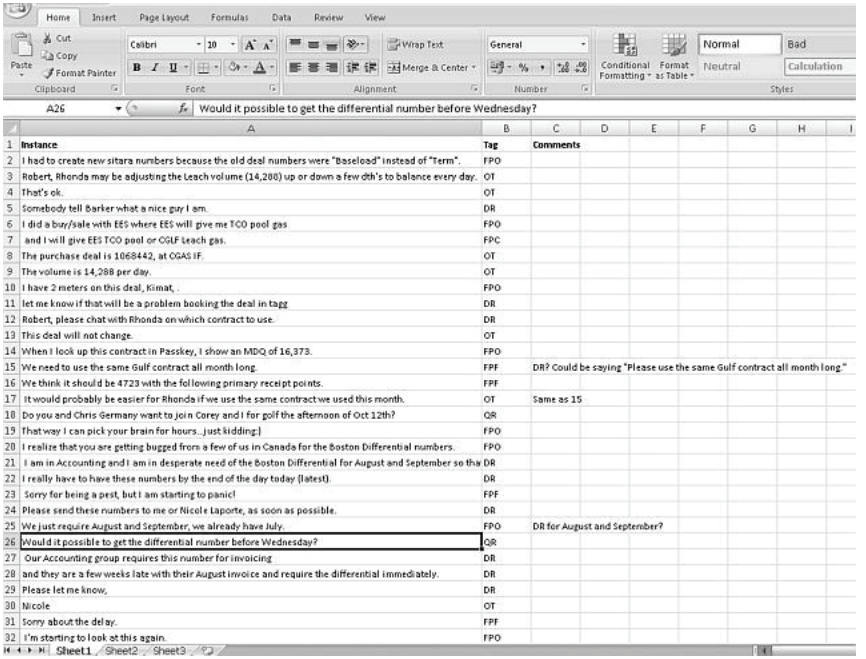
In this section, we describe the annotation procedure. We first provide a general overview of the process, and then proceed to discuss the criteria used to assign individual utterances to the different categories by discussing in detail several instances which were found difficult to annotate.

5.1 Annotation tools and procedure

A key requirement of the annotation procedure is ease of implementation, meaning that it should be relatively easy both to train new annotators and to carry out the annotation itself. By simplifying the technical aspects, we can draw on the skills of a number of linguistics graduate students, regardless of their technical expertise. Two annotators (native English speakers enrolled in linguistics PhD programs) tag each utterance – further discussion of their agreement levels is below. Each utterance is to receive one tag only; the implications of this decision are discussed below and in Section 6.

There is only one level of annotation, namely the speech act category. The data is stored as plain text, for ease of analysis by NLP and command line tools, but is presented to the annotators in an Excel spreadsheet, one line per cell, with the tag to be inserted in the adjacent column, and any relevant comments in the column adjacent to that. This makes it easy to distribute the tagging materials, carry out annotation offline, collate the annotations and filter out instances as required. Each speech act category has an abbreviated tag, designed to be mnemonic, as listed in Table 1 above. There is also a 'comments' column for the annotators to remark on any difficulties experienced in assigning a tag, or to note any uncertainty in the choice of tag assigned on the instance. This has proved particularly valuable for the research stage, both as an easy way to identify interesting or liminal cases for study, and as a way to gather further insights

into what human readers rely on as cues to speech act identification. Figure 1 shows a screenshot of some annotated data, including examples comments. In the initial phase of the annotation procedure, feedback was sought from the annotators regarding the methods chosen for the annotation and the ease with which it was possible to remember the tags while annotating. All three annotators responded positively, enabling us to proceed with a high degree of confidence in the tagset and annotation procedure.



	Instance	Tag	Comments
2	I had to create new sitara numbers because the old deal numbers were "Baseload" instead of "Term".	FPO	
3	Robert, Rhonda may be adjusting the Leach volume (14,288) up or down a few dth's to balance every day.	OT	
4	That's ok.	OT	
5	Somebody tell Barker what a nice guy I am.	DR	
6	I did a buy/sale with EES where EES will give me TCO pool gas	FPO	
7	and I will give EES TCO pool or CGLF Leach gas.	FPC	
8	The purchase deal is 1068442, at CGAS IF.	OT	
9	The volume is 14,288 per day.	OT	
10	I have 2 meters on this deal, Kimat, .	FPO	
11	let me know if that will be a problem booking the deal in tagg	DR	
12	Robert, please chat with Rhonda on which contract to use.	DR	
13	This deal will not change.	OT	
14	When I look up this contract in Passkey, I show an MDQ of 16,373.	FPO	
15	We need to use the same Gulf contract all month long.	FPF	DR? Could be saying "Please use the same Gulf contract all month long."
16	We think it should be 4723 with the following primary receipt points.	FPF	
17	It would probably be easier for Rhonda if we use the same contract we used this month.	OT	Same as 15
18	Do you and Chris Germany want to join Corey and I for golf the afternoon of Oct 12th?	QR	
19	That way I can pick your brain for hours...just kidding:)	FPO	
20	I realize that you are getting bugged from a few of us in Canada for the Boston Differential numbers.	FPO	
21	I am in Accounting and I am in desperate need of the Boston Differential for August and September so tha	DR	
22	I really have to have these numbers by the end of the day today (latest).	DR	
23	Sorry for being a pest, but I am starting to panic!	FPF	
24	Please send these numbers to me or Nicole Laporte, as soon as possible.	DR	
25	We just require August and September, we already have July.	FPO	DR for August and September?
26	Would it possible to get the differential number before Wednesday?	QR	
27	Our Accounting group requires this number for invoicing	DR	
28	and they are a few weeks late with their August invoice and require the differential immediately.	DR	
29	Please let me know,	DR	
30	Nicole	OT	
31	Sorry about the delay.	FPF	
32	I'm starting to look at this again.	FPO	

Figure 1: Screenshot of annotation process

The training material was also intentionally simple, to enable annotators to start the task quickly. Prospective annotators were given a short introduction to the project with some general directions, as well as three files to assist them in the task: a three page description of all the categories, including three sample annotated emails; a summary 'tag table' with two or three examples for each category; and a flowchart to assist in streamlining decisions. These materials are found in the Appendices. They also received a sample set of 100 lines to anno-

tate, both to familiarise themselves with the project and to identify any other potential issues.

While it was relatively easy to decide the categories of interest, creating the actual description of each category to assist the annotators was a greater challenge, as there are few resources available from which to draw inspiration. Theoretical descriptions of speech acts in the linguistics literature (for example Levinson 1983; Cruse 2000; Jaszczolt 2002) were not found to be particularly useful guidelines in creating annotation instructions, as they make generic reference to lexical items or verbal mood and tense. We therefore employed a common sense approach, whereby the different categories are described with reference to both surface features (e.g. subject, presence of particular phrases) and to more subtle, contextual features of the text (e.g. realising from the context that a request is being made). The guidelines also include examples of instances which could be assigned to more than one category, noting the preferred tag for these. Overall, neither the project leader nor the annotators feel there are any problems with the setup of the task.

Alternatively, we could have proceeded in an entirely data-driven way, selecting some features for the utterances, then applying a clustering algorithm to see how they automatically grouped themselves on an empirical basis. This approach is planned for future research, but in the first instance, to better align with previous research in linguistics, it was deemed more appropriate to assign the data to pre-determined categories, and use NLP tools to verify how homogeneous the categories are.

At all times during the annotation procedure, the annotators have access to the entire email text. This can be helpful in cases where a particular utterance could be interpreted in more than one way, as the annotators can refer to contextual clues for disambiguation. On the other hand, if one is interested in carrying out a category-by-category analysis, in which only decontextualised instances are available, it can lead to a certain amount of annotator confusion if the role of a particular instance is not obvious. One of the challenges of this kind of pragmatic research is how to include the role of wider context in examining primarily linguistic features.

As mentioned previously, the annotation process so far has yielded approximately 20,700 tagged utterances. The distribution of the different categories is show in Table 2: we can see it is somewhat uneven, with statements (FPO + OT) making up almost sixty percent (58%) of the data. All the utterances are double-annotated; one of the annotators was constant throughout the process. On average, annotators achieved rates of up to 500–550 utterances tagged per hour, depending on the complexity of the utterance and the detail given in the com-

ments. In Section 5.2, we discuss some of the problems related to annotator disagreement.

Table 2: Distribution of speech act categories in the data

Speech act	**Percentage of total**
Other statements (OT)	41.4%
First person other (FPO)	16.5%
Direct request (DR)	13.6%
First person commitment (FPC)	10.5%
First person expression of feeling (FPF)	7.9%
Open question (QQ)	7.3%
Question-request (QR)	2.7%

5.2 *Contentious cases*

Pragmatic interpretation is often a subjective task, so it is not uncommon to have a high frequency of disagreement in the annotation. In our data, average agreement between the annotators varies greatly between the two annotator pairs, with one pair having average agreement of seventy-one percent (weighted kappa 0.78) and the other pair eighty-two percent (weighted kappa 0.87). Further discussions and analysis are needed to understand the possible causes of this divergence, which is particularly striking since the annotators who obtain higher agreement differ in gender, nationality, and location.

Where annotators disagree, there are several attempts to achieve a reconciliation. In some cases, it is simply a clear mistake on the part of the annotator, such as a typo or incorrect selection (for instance due to autocomplete), and can be easily rectified. In other cases, where both tags suggested may, in certain interpretations, be plausible, we attempt to come to an agreement through discussion. However, this is not always possible, especially in cases, as we will see below, where the function of the utterance is highly ambiguous and it is genuinely impossible for annotators to decide. In the interest of creating a gold standard dataset for research purposes, these contentious cases are not included in the corpus for analysis, but they are retained separately as a dataset of 'interesting cases' both for linguistic reference and for further testing of the robustness of the NLP tagger (described in Section 7).

Our approach to annotation disagreements, with case-by-case discussions, has the disadvantage of increasing the duration of the annotation process, and is open to a certain amount of human bias. If one had three annotations instead of two for each instance, it would be easier and faster to resolve these disagreements by relying on majority tags, but additional annotators would add to the project's time and costs. It is, however, something to be advocated where resources permit.

From the point of view of pragmatics research, however, the instances where annotators disagree or express uncertainty prove highly informative. These disagreements highlight recurring problems with the task and with the issue of speech act classification more generally. The problems encountered during the annotation procedure can be summarised by noting that we are taking a formal approach to functional, pragmatic categories. As the annotation guidelines show, there is often reference to surface features in determining which category to assign an utterance to, but the formal aspects can contrast sharply with the intended function of the utterance. Furthermore, the speech acts often had multiple functions and it is difficult to determine a single category for these.

In addition to the issues mentioned above, there are also purely formal difficulties relating to the nature of email writing, especially in fast-moving workplace environments where correct prose is not required for routine casual communication. In these emails, pronoun-dropping was fairly common, which made it difficult to tell exactly what function the clause was performing, as in this example: *Also presume to message the letter from SK to the Senators and our solution*. In other cases, the content of the utterance referred to very domain-specific topics which eluded the annotators' understanding, and made it impossible to categorise the utterance.

We now look at some examples of challenging utterances more closely, to illustrate the types of problems that arise. The issues discussed, and the possible solutions outlined, are of a general nature and can be applied to the wider context of speech act research.

5.2.1 First person categories

The first person categories were a particularly rich source of discussion for all annotators, and led to confusion and uncertainty in various ways. One confusion pair which emerged from the comparison of the annotators' tags, although not one on which any of them remarked explicitly, is the distinction between first person statements (FPO) and commitments (FPC). From the disagreement patterns, we can see that it was not always easy to differentiate between the two, a problem illustrated in sentences such as the following:

(2) We continue to consider every option available to us.

(3) We are taking steps to be prepared to isolate the TMS system.

Formally, we can see that these lines can be assigned to either the FPC or FPO category. Although they do not have the form of the most prototypical commitment (such as *I will do X, I will be doing X*), they could be interpreted as an undertaking on the part of the speaker to perform the action described (*consider options, isolate the system*). It is possible that the commitment function of utterances such as these may become clearer in the context of the email in its entirety, as for example when a number of tasks are assigned to various individuals or groups within the same email. On the other hand, it could just be one of a series of facts being imparted, such as a description of an ongoing event or project. Without knowledge of the wider extra-textual context, that is to say, of the business environment and its associated discourse practices, it is very difficult to know how to interpret utterances such as these.

However, such potential sources of confusion do not undermine our decision to distinguish between FPOs and FPCs. Traditionally, speech act classifications distinguish between representatives and commissives, and there are clear areas where this distinction makes sense. For instance, it would be hard to claim that statements such as *I am the author of this article* or *We work in a university* fulfil the same function as statements such as *We will finish the article by the deadline* or *I will be presenting at this conference*. Examples such as the above fall into a grey area between the two, and understanding how to negotiate the relationship between form and function is a key aspect of our research.

The third class of first person statements, 'first person feeling' (expressives, FPF), was also a major point of discussion for several reasons. A particular trigger for confusion was the different functions associated with the phrase *I think*. For example, one of the annotators noted a consistent difficulty in distinguishing FPO from FPF in statements such as *I think you need this information*. The tagging guidelines indicated that these kinds of statements should be tagged as FPO, as the first-person construction is primarily mitigating the statement. However, at least one of the annotators observed a difficulty in clearly distinguishing between instances where the phrase was fulfilling a 'mitigating' role and those where it was 'offering an opinion'. There are many kinds of mitigating phrases of this kind that led to these interpretation problems. Another annotator remarked upon the use of expressions that perform negative politeness/mitigation, such as *I don't know about you guys, but…* (followed by a request). This is, formally, a fairly straightforward case of FPF – explicitly containing the cognitive verb *know*. However, the pragmatic function of this utterance seems to be to

perform negative politeness work, offering the others the chance to disagree with the sender. In this sense, the preface *I don't know about you guys* functions as part of an indirect request (QR), attempting to elicit information from the addressees as to whether they have also experienced this problem. While the surface form of examples like these may suggest an interpretation as FPF, it is possible that such utterances possess a variety of additional functions that cannot be identified by attention to formal features alone.

In other cases involving embedded clauses in the expressives category, the discussion centred on whether to give priority to surface form or pragmatic function, as in the following:

(4) I think the results are due out today.

(5) I know Mark is working on the report.

Here, although the grammatical form is that of a 'first person feeling' speech act, the main message of the sentence is actually the objective, third person statement in the embedded clause; the first person statement is mainly there as a hedge or qualifying phrase. They are clearly fulfilling a different role from 'pure' expressions of feeling such as *I think I am too tired for the staff party* or *I know we'll be very impressive at our presentation*, which report subjective facts.

The disjuncture between form and function for FPF and OT also occurs in the opposite direction, where two of the annotators have commented on the limitations of the scheme with regard to utterances which are in the third person but express an emotional response, such as:

(6) It's fun and games every day.

(7) That would be great.

Here, the utterances have the illocutionary force of an FPF, as they express the evaluative opinions of the email sender, but are lacking the definitive linguistic features of a first-person expressive. The first-person pronoun, one of the central features leading to membership of the FPF category, is absent, and it is only the metaphorical (and colloquial) expression of *fun and games*, or the adjective *great*, that give a sense that these are opinions rather than general observations belonging to the OT category. Despite this, these utterances do not qualify as FPF on formal grounds, so, according to the annotation guidelines, they must in fact be tagged OT.

Surface lexical features appear to be a strong trigger for the FPF class, according to two of the annotators, for example in the case of other cognitive verbs such as *would rather*, *expect*, *recommend*, and *prefer*. Consider the following:

(8) At this juncture I would recommend against having Ken participate.

(9) I would recommend they do not sign.

These could be interpreted as statements of the writers' feelings about the situations indicated, or, if the writer were known to have the authority to influence or dictate the behaviour of the referents or addressees, they could be said to possess the illocutionary function of a mitigated request or advisement. In most cases, the presence of those verbs leads to the choice of FPF.

Currently, the annotation guidelines tend to steer annotators towards relying on surface cues in assigning speech act categories to utterances. As this discussion shows, however, there is a risk that this approach is too broad and cannot account for more subtle distinctions in the way language is used. On the other hand, this form-oriented approach may have the advantage of easing the cognitive load for the annotators, limiting the amount of time they have to spend on each utterance. Clearly, there is a conflict between the desire to create an information-rich and accurate annotation procedure and the need to minimize the burden placed on annotators. These competing imperatives must be subject to further careful assessment if an acceptable balance is to be found.

5.2.2 Other categories

Overall, the annotators reported no particular problems with the 'other' (OT) category, except for the OT/FPF confusion discussed above. One annotator also remarked on the possible confusion between the two types of question categories, noting that occasionally there seemed to be overlap between a pure question asking for information, and a question used as a request, for example in *Would there be a good time to visit tomorrow?*. In cases like these, there seemed to be a consistent difference of opinion between annotators. While one tagged these as straightforward questions (QQ), the other felt that such cases could either be interpreted as QQs or as a polite phrasing of an indirect request (QR), for instance, *Can I visit you sometime tomorrow?*. However, the other annotators did not remark on this difficulty. This is a difficult case to arbitrate because it is a very good example of the form-function distinction, which we will discuss at greater length below. Arguably, in these cases, where we are missing the relevant contextual information that would allow us to fully understand the illocutionary force of the act, we should focus only on the form (as indeed some of the annotators do) and tag these as straight questions. The risk is, however, that in doing so we may lose useful examples of the wide range of strategies adopted for indirect requests.

A further challenge arises from the fact that questions are not the only form requests can take. One of the annotators observed that it was often difficult to distinguish directives (DR), as these were frequently linguistically mitigated, as in the following: *At some point, the document must be finished*. This could either be classed as OT or DR, depending on whether it is read as a general description of a state of affairs or as a specific action that the sender wants the addressee to perform. Without key information about the relationship between the sender and the addressee, it is impossible to decide with any degree of confidence whether this should be OT or DR. If the sender were known to be in a position of authority, it would at least be possible to interpret this line as a directive to the addressee (albeit mitigated); if there is no power difference between sender and addressee (or if the addressee is the more powerful interlocutor) we would be more likely to assign the utterance to the OT category. This issue of sender/addressee power relations also arises with first person statements. Consider for example the statement *I need the figures for the report*: this can be read as a request if it is said to someone more junior than the speaker, while if the hearer is a peer, or a senior, it could be simply the acknowledgement of a fact, or even a commitment on the part of the speaker to find those figures.

6 The form-function disjuncture

The form-function disjuncture for the FPF-OT pair is a good example for the discussion of how to address this issue more generally in a pragmatic annotation task, as it would be sensible to decide a priori for the task which one to prioritise in assigning categories to the data. One might argue that function should be privileged in this case because we are interested in learning about speech act functional categories, not forms. Including these statements in one category rather than the other is misleading and misrepresentative, because we want to learn about *all* the different ways in which this function is expressed. On the other hand, it is not always easy to assess where the illocutionary meanings (for example the hedging functions of the first person statements) begin and their locutionary, literal, expressive meaning ends, and we might be overestimating their representative function. Ultimately, both positions are tenable, and we argue that the deciding factor should be the underlying research aim: if we are interested in a clear picture of 'pure' expressives, we might want to avoid including these statements. If we are interested in gathering together all the types of statements that individuals make about themselves, then we should include them. If, finally, we are interested in gaining a clear picture of all the ways in which statements of fact are represented, we ought to tag these as OTs instead. From a methodologi-

cal point of view, it is worth noting that deciding to operate this distinction can slow down the annotation process, as, rather than semi-automatically assuming that all statements beginning with *I think/know…* are FPFs, reflection is needed for each one. It would also be necessary to assess how these changes would affect the performance of our automated tagger. Further solutions are explored in more detail in Sections 6.1 and 6.2.

More importantly, establishing a position on the form-function distinction has implications for the project as a whole, as it would be advisable to maintain a coherent position on this issue across all categories. For example, if we were to decide that content and function have priority over surface form, we ought to collapse the request categories into a single one including both direct and indirect requests. This would then make it difficult to engage in close study of the different types of request, which is an area of great interest in the literature.

The previous discussion has highlighted some theoretical and practical concerns relating to pragmatic interpretation at the level of form and function. More generally, an important question to ask, given the dependence of pragmatic interpretation on context, is how far the annotation process can reasonably proceed without access to the relevant extra-linguistic information. The process seems to highlight the complicated relationship between the meanings of the words used and the context of the utterance. In particular, as illustrated in a number of examples above, there are occasions where additional knowledge of the interpersonal context – the relative power of interlocutors, the nature of their relationship and relevant shared histories – may be helpful in determining the function of the sentence in question.

Unfortunately such information is not available to us with this particular corpus of Enron emails, since, as previously noted, the corpus was not designed to aid linguistic research (though it is theoretically possible with some effort to reconstruct at least some relationships). An open question that therefore remains for pragmatic annotation is how to know when to acknowledge that no further interpretations and assumptions can be made about the data, and an upper bound has been reached.

6.1 Addressing the multifunctionality problem

In many cases outlined above, the confusion between form and function might have been resolved with full access to the relevant context. However, sometimes the utterances do seem to genuinely function as more than one speech act. For example, *Can't wait to catch up* is a first person statement which serves as an expression of the writer's emotional response at the prospect of meeting up with the addressee, and can therefore be tagged FPF. However, it can also be inter-

preted as functioning to reassert the writer's commitment to a possible or forthcoming future meeting, and can therefore also be tagged FPC. Such multifunctional speech acts are, of course, not rare in language and our reliance on a tagset that does not take into account that this multi-functionality necessarily forces the tagger, be it human or machine, to make an arbitrary decision. Adopting a more flexible scheme would allow us to treat these cases appropriately, and also address the form-function disjunctures. What form should a more flexible scheme take?

One possibility would be to have every utterance receive two tags, one for its locutionary and one for its illocutionary meaning, to capture the difference between form and function. In practice, each utterance would have two columns for tags, one for each meaning, where the tags may or may not coincide. However, this would significantly complicate the annotation procedure, placing a greater burden on the annotator, who would be required to make twice as many decisions. Furthermore, this dual tag approach would have no impact on the problem of utterances that are open to more than one illocutionary interpretation. On the other hand, it would provide a straightforward way of gathering information about the most common combinations of form and function. A more simple solution would be to allow utterances to receive more than one category tag when the annotator could not decide between them, or felt there was more than one function carried out concurrently, without specific reference to form and function. This would require less time but might lead to the annotators actually reducing their effort, and over-relying on multiple tags when faced with a challenging instance. Its advantage would be to allow easy identification of the kinds of utterances that can have multiple interpretations.

Another possibility would be to include a special 'disjuncture' flag for problematic utterances, such as the FPF-OT pairs discussed above. In this case, we would not be told what the other possible tag is, but merely that a disjuncture has been interpreted, for example by using a special letter flagging this. This approach would have the advantage of greater annotation speed and lighter cognitive load, but at the expense of informativeness. The three possible solutions discussed all have advantages and drawbacks, though they all ultimately point to the need for a more sophisticated approach to speech act tagging that goes beyond the one line, one tag method. We are now planning to run some pilot annotation tasks exploring the use of all three alternatives to clearly assess which responds best to the needs of both annotators and researchers, as a further contribution to the research community engaged in these tasks.

6.2 Multi-utterance problems

Many of the problems highlighted above derive from local issues about the mismatch between form and function in particular utterances. However, the strategies adopted by the annotators to assess the correct tag lead to further problems regarding pragmatic annotation: is it a local or a global task? So far we have been discussing tagging and categories as a very atomistic entity, and in fact the guidelines make no reference to whether the annotators should read messages globally or proceed line by line, ignoring surrounding text. In practice, the latter approach is impossible, because the set up of the annotation task is such that the entire email is there for the annotators to read, and it would be unrealistic to do otherwise. Indeed, focusing on decontextualised utterances would likely result in the assignment of tags based solely or primarily on the locutionary aspect of speech acts, because decontextualised speech acts are very difficult to interpret. The problem then becomes a multisentential one: can a speech act span across more than one sentence? The following example is given:

(10) Jeff – if the lawyers can't, I'm sure we can ask M to get the filing. Let me know.

Viewed in isolation, the first utterance would be an FPF (i.e. epistemic *I'm sure*). However, (re)viewed in the context of the DR *Let me know* that follows it, we could read the first utterance as another directive – one that performs negative politeness work by minimising the imposition on the hearer and avoiding a direct request. This is a valid concern since it is possible to find texts where the illocutionary effect of a request comes not from a single sentence but from the entire sequence of utterances. Such an analysis is, at present, beyond the scope of our work, though we acknowledge the important role of utterance sequences in speech act interpretation. This information is also very valuable for non-native language speakers interested in learning the appropriate speech act sequences in English. Therefore, the analysis of discourse structure will be included in the research carried out on the annotated data, examining issues such as the placement of requests and related supporting utterances within the text of the email.

In summary, the main problems encountered by the annotators relate to the fact that lexicogrammatical categories were being used to identify pragmatic phenomena, while the tag-set did not allow for the multifunctionality that was often observed in the utterances that comprise the data set. Furthermore, the absence of contextual information, in particular that pertaining to the relationships between writer and addressee(s), their degree of intimacy, and relative status, sometimes made it difficult to determine the role of particular utterances.

7 Using the data

The data so far annotated has been used to further work in both computational linguistics and applied linguistics, through the development of two tools, SPADE (SPeech Act Descriptor for English) and SPATE (SPeech Act Tagger for English). This aspect of our project falls within the domain of 'cue-based' or probabilistic computational models for speech act interpretation, which are described in detail by Jurafsky (2004). In these models, the task is to identify, from the surface form of the sentence, the relevant cues to the speaker's intentions. Typical cues include lexicon, collocations, syntax, prosody, and conversational structure (Jurafsky 2004: 588); our model includes all of these except, of course, the prosodic aspect, since we are not working with spoken data. This section provides only a brief introduction to the types of research that arise from this corpus; further details can be found in De Felice (2011b) and De Felice (forthcoming 2013).

SPADE is a natural language processing tool focused on the linguistic interpretation of speech acts. Its main goal is to proceed from the annotated data to the creation of detailed, multi-level pragmatic descriptions of different speech act categories. It is designed to work with the output of the C&C toolkit (Curran *et al.* 2007), which consists of a set of applications including a morphological analyzer (Minnen *et al.* 2001), a part-of-speech tagger, the supertagger described earlier, a CCG parser (Clark and Curran 2007), and a named entity recognizer (Curran and Clark 2003). Each utterance is analysed and parsed by the C&C tools, and from the output SPADE extracts information about its lexicon, grammar, and syntactic structure which aids the linguistic description of the speech act. Figure 2 shows the full list of features extracted by the tool. The presence of particular n-grams is also noted, though the implementation of this feature requires further analysis. Figure 3 shows a sample output from SPADE: among other features, the tool has identified the absence of modal verbs and adjectives, the presence of two direct objects and of the key unigram *appreciate*, the use of a first person subject, and the declarative nature of the sentence ('S[dcl]').

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

Figure 2: Features extracted by SPADE

```
I appreciate your continued patience and cooperation.  [FPF]
['Modal': 'no', 'LastWord': 'cooperation', 'Object': 'noun',
'Punct': '.', 'Object2': 'noun', 'PredicativeAdj': 'none',
'VerbTag': 'VBP', 'FirstWord': 'I', 'SubjectIs': 'Firstperson',
'SentenceType': 'S[dcl]', 'HasUnigram_appreciate': 'yes',
'Subject': 'pronoun']
```

Figure 3: Sample output of SPADE

This analytical framework extends what is possible with common concordance tools by including grammatical and syntactic annotation which allow for more sophisticated queries. The availability of data thus processed enables the investigation into the linguistic properties of speech acts. The database-like format of the data provides an efficient way of identifying and searching for significant patterns and prominent features. For example, to query the data about which speech act category uses transitive verbs most often, we could look at the frequency of the feature 'Object' with values 'yes' and 'none' for each class, and discover that requests and commitments have the highest proportion of transitive verbs, while they are very low in statements. Other items of interest include the role of proper nouns, the use of adverbs in different speech acts, and typical

subject-verb combinations, among many others. This information is also of benefit to the applied linguistics and language teaching community, especially where it is possible to carry out comparisons with similar non-native speaker data.

The SPADE output is also the essential and necessary component of the SPATE tool, which is currently in its second stage of development (for details of the first version, based on non-native language, see De Felice and Deane 2012). The tagger uses a maximum entropy machine learning classifier for automated speech act tagging, trained on the features extracted by SPADE. In other words, it learns to associate particular combinations of features to a given speech act category, so that it can correctly assign a speech act category to a novel, previously unseen instance. Preliminary experiments run on small subsets of the data have shown that the tagger currently achieves around seventy-five percent accuracy (precision 74.5%, recall 68%), but as more annotated data becomes available for training and testing, these figures are expected to improve. It performs particularly well on the OT, QQ, and DR classes and, similarly to the human annotators, obtains lower figures on the first person categories, especially FPF.

The SPATE tool is useful for the corpus linguistics community as an additional means of exploring corpus data. Furthermore, by examining its error patterns, and the items it finds particularly challenging to classify correctly, we can obtain further insights on what constitute non-prototypical cases of each category. For example, this set of incorrectly classified QQs draws our attention to the fact that not all questions using modal verbs are requests, a fact that is of interest in particular to language learners:

(11a) Could the gift recipient select the menu items?

(11b) Should we attach the first round of questions?

(11c) Can these really all be receipt imbalances?

These kinds of examples, together with the discussion of the annotators' comments and disagreements discussed above, and the patterns found in the data, all contribute to raise questions regarding speech act categorisation, which remains open to debate. Archer *et al*. note that "Pragmatic interpretations, leading to the implementation of a functional tag (e.g. a speech act), require a complex synthesis/understanding of contextual information that is currently beyond the means of a computer" (Archer *et al*. 2008: 634). Indeed, as we have seen in the discussion above, sometimes this interpretation also eludes humans. On the other hand, our results suggest that for a large number of utterances it is possible to carry out automated tagging with a reasonable chance of success.

8 Conclusions

In concluding their article, Archer *et al*. propose that annotation guidelines should consider "the need to devise an annotation scheme in relation to one's research goals, the need to be systematic enough to ensure replicabililty (and, by so doing, ensure its usefulness to others), the need to balance delicacy of categorisation with the ability to fill categories with a statistically meaningful quantity of members, and so on" (Archer *et al*. 2008: 638). The annotation project described in this paper responds to these requests. Our annotation scheme supports our research endeavour of gaining an understanding of speech act use in business communication without being too task-specific; the categories used ensure that all categories are large enough for meaningful analysis; and by publishing the guidelines used, together with unresolved issues, we hope to stimulate some discussion about the reusability of our classification scheme.

We have shown that, despite the constraints of context-dependent pragmatic interpretation, it is possible to implement a pragmatic annotation scheme which can yield valuable insights into the communicative processes of email in the business domain. We plan to make the corpus data freely available to invite further engagement with both the methodological and the linguistic issues that arise from the research. Our work to date has also demonstrated some practical applications of the annotated data, contributing to the fields of linguistics and NLP.

Open questions remain, in particular regarding the role of contextual information, and the best way to handle utterances of ambiguous interpretation. There is further work to be done in establishing a classification scheme that avoids some of the problems discussed in this paper. O'Keeffe *et al*. note that "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: 97). Our classification scheme is modelled closely on traditional speech act theory, but perhaps our findings will point us towards a revised model that mirrors more closely how we *really* 'do things with words'.

Acknowledgements

This work was carried out while Rachele De Felice was supported by a Leverhulme Early Career Fellowship at the University of Nottingham; this research was undertaken by the Fellowship holder and not on behalf of the Leverhulme Trust. We gratefully acknowledge the support of a research award from the University of Nottingham Research and Knowledge Transfer Board. We thank Svenja Adolphs, Ron Carter, and Paul Deane for insightful conversations during

the development of this project, and the anonymous reviewer for helpful comments.

Notes

1. The supertagger is a preprocessing step in statistical parsing using CCG. It assigns to each word a set of the most probable CCG lexical categories given the context, describing its syntactic behaviour before determining the sentence's full syntactic structure. Lexical categories can be either basic (such as NP, VP) or complex (combinations of argument and result categories, such as VP/NP, a verbal category which takes an NP argument and results in a VP).

References

Archer, Dawn. 2005. *Questions and answers in the English courtroom (1640–1760): A sociopragmatic analysis*. Amsterdam: Benjamins.

Archer, Dawn, Jonathan Culpeper and Matthew Davies. 2008. Pragmatic annotation. In A. Lüdeling and M. Kytö (eds.). *Corpus linguistics: An international handbook.* Vol. 1, 613– 642. Berlin: Walter de Gruyter.

Austin, J. L. 1962. *How to do things with words*. Oxford: Clarendon Press.

Berry, Michael, Murray Browne and Ben Signer. 2007. *2001 Topic Annotated Enron Email Data Set*. Philadelphia: Linguistic Data Consortium.

Bjorge, Anne Kari. 2007. Power distance in English lingua franca email communication. *International Journal of Applied Linguistics* 17 (1): 60–80.

Carvalho, Vitor and William Cohen. 2005. On the collective classification of email "speech acts". *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 345–352. New York: Association for Computing Machinery.

Carvalho, Vitor and William Cohen. 2006. Improving email speech act analysis via n-gram selection. *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech*, 35–41. Association for Computational Linguistics.

Clark, Stephen and James Curran. 2004. The importance of supertagging for wide-coverage CCG parsing. *Proceedings of the 20th International Conference on Computational Linguistics (COLING),* 282–288.

Clark, Stephen and James Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics* 33 (4): 493–552.

Core, Mark and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. *Proceedings of the Working notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines,* 28–35.

Cruse, Alan. 2000. *Meaning in language: An introduction to semantics and pragmatics*. Oxford: Oxford University Press.

Curran, James and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. *Proceedings of the 7th Conference on Natural Language Learning* (*CoNLL),* 164– 167. Association for Computational Linguistics.

Curran, James, Stephen Clark and Johan Bos. 2007. Linguistically motivated large-scale NLP with C&C and Boxer. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions,* 33–36. Association for Computational Linguistics.

De Felice, Rachele. 2011a. Language at work: Native and non-native speech acts in Business English. Paper presented at the Joint Conference of the *BAAL Intercultural Communication Special Interest Group* and *The Annual Bloomsbury Round Table*.

De Felice, Rachele. 2011b. Pragmatic profiling of business corpora: Speech act tagging. Paper presented at the *32 ICAME* conference.

De Felice, Rachele. Forthcoming 2013. A corpus-based classification of commitments in Business English. In *Yearbook of Corpus Linguistics and Pragmatics* 1.

De Felice, Rachele and Paul Deane. 2012. *Identifying speech acts in emails: Toward automated scoring of the TOEIC(r) email task*. Princeton, NJ: ETS.

Geertzen, Jeroen, Volha Petukhova and Harry Bunt. 2007. A multidimensional approach to utterance segmentation and dialogue act classification. *Proceedings of the 8th SIGDIAL Workshop on Discourse and Dialogue,* 140–149.

Georgila, Kalliroi, Oliver Lemon, James Henderson and Johanna Moore. 2009. Automatic annotation of context and speech acts for dialogue corpora. *Natural Language Engineering* 15 (3): 315–353.

Gimenez, Julio. 2006. Embedded business emails: meeting new demands in international business communication. *English for Specific Purposes* 25 (2): 154–172.

Goldstein, Jade and Roberta Sabin. 2006. Using speech acts to categorize email and identify email genres. *Proceedings of the the Hawaii International Conference on System Sciences*.

Ho, Victor. 2010. Contrasting identities through request e-mail discourse. *Journal of Pragmatics* 42: 2253–2261.

Hockenmaier, Julia and Mark Steedman. 2007. CCGBank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics* 33 (3): 355–396.

Jaszczolt, Kasia 2002. *Semantics and pragmatics: Meaning in language and discourse*. London: Longman.

Jensen, Astrid. 2009. Discourse strategies in professional e-mail negotiation: A case study. *English for Specific Purposes* 28 (1): 4–18.

Jurafsky, Dan. 2004. Pragmatics and computational linguistics. In L. Horn and G. Ward (eds.). *The handbook of pragmatics*, 578–604. Oxford: Blackwell.

Kallen, Jeffrey and John M. Kirk. 2012. *SPICE-Ireland: A user's guide*. Belfast: Cló Ollscoil na Banríona.

Klimt, Bryan and Yiming Yang. 2004. The Enron corpus: A new dataset for email classification research. *Proceedings of the European Conference on Machine Learning (ECML),* 217–226.

Lampert, Andrew Robert Dale and Cecile Paris. 2008. The nature of requests and commitments in email messages. *Proceedings of the AAAI Workshop on Enhanced Messaging*, 42–47.

Lampert, Andrew, Robert Dale and Cecile Paris. 2010. Detecting emails containing requests for action. *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL),* 984–992. Association for Computational Linguistics.

Lendvai, Piroska and Jeroen Geertzen. 2007. Token-based chunking of turn-internal dialogue act sequences. *Proceedings of the 8th SIGDIAL Workshop on Discourse and Dialogue,* 174–181.

Leuski, Anton. 2005. Context features in email archives. *Proceedings of the 28th International SIGIR Conference on Research and Development in Information Retrieval, Workshop on Information Retrieval in Context (ACM SIGIR IRiX)*, 54–56.

Levinson, Stephen. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

Maynard, Carson and Sheryl Leicher. 2006. Pragmatic annotation of an academic spoken corpus for pedagogical purposes. In E. Fitzpatrick (ed.). *Corpus Linguistics beyond the word: Corpus research from phrase to discourse*, 107–116. Amsterdam: Rodopi.

McEnery, Tony and Andrew Wilson. 1996. *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Minnen, Guido, John Carroll and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering* 7 (3): 207–223.

Newton, Jonathan and Ewa Kusmierczyk. 2011. Teaching second languages for the workplace. *Annual Review of Applied Linguistics* 31: 74–92.

O'Keeffe, Anne, Brian Clancy and Svenja Adolphs. 2011. *Introducing pragmatics in use*. London: Routledge.

Stiles, William. 1992. *Describing Talk: A taxonomy of verbal response modes*. Thousand Oaks, CA: Sage.

Stolcke, Andreas, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Dan Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics* 26 (3): 339–371.

Styler, Will. 2011. *The EnronSent Corpus*. Boulder, CO: University of Colorado at Boulder Institute of Cognitive Science.

Ulrich, Jan, Gabriel Murray and Giuseppe Carenini. 2008. A publicly available annotated corpus for supervised email summarization. *Proceedings of the AAAI Workshop on Enhanced Messaging,* 77–82.

Appendix A: Annotation guidelines

Speech act annotation – descriptions and examples of each category
Training version – March 10, 2011

Tags:

DR	QR	QQ
FPC	FPF	FPO
OT	EX	

The text is already broken up into the units that require tagging. One unit = one tag.

For sentences containing conditionals, generally the emphasis should be on the content of the main clause and the tag should reflect that. So, in a sentence such as *If I send you the files now, can you reply by tomorrow?*, the tag should reflect the speech act represented by *can you reply by tomorrow.*

The annotation scheme tries to take into account both form and function of the act by using the same letter when these coincide across different types of act. For example, **R** indicates a request, so it appears in the two tags which refer to two kinds of request. **FP** refers to first person statements, of which there are three kinds, as the tagset shows. I have also tried to make them somewhat easy to remember.

Questions and Requests – Q and R
These are described together since they overlap to some extent.
For something to be tagged as having a question form, it can either be a direct question, ending with a question mark, or be embedded in a declarative clause (e.g. *I want to know what time it is* or *I wonder how I much I can get for this job.*).

Pure questions – **QQ** – are those which are a genuine request for information which can be obtained without the hearer having to take special action. For example: *What's your name? What time is the meeting?*

Requests can be formed as questions or more direct orders; their defining characteristic is that they request that the hearer do something, they attempt to affect the hearer's behaviour, whether by requiring an action (*Please send me the files. I'd like to ask you to phone the client for me.*) or a change in their mental state

(*Don't worry about the meeting. You should feel more positive about the interview.*). Second person pronouns feature heavily here.

If they are formulated as imperatives or statements, such as all the examples in the previous paragraphs, the tag is **DR** – for 'Direct Request'.

If they are phrased as questions, the tag is **QR** – for 'Question Request'. Some examples of this are *Could you send me the files? Would you be able to drive us to the office? I was wondering if you could give me a call.* These can be very tricky, as sometimes one will use very roundabout ways to issue requests due to extreme politeness. In general, if the answer to the question requires any sort of action on the part of the recipient, it is to be considered a request. An example could be the question *Will I be doing the same work as before?*, which could be interpreted either as a straightforward question with a yes or no answer, or as a very indirect way of saying "Please tell me what I need to do". In the past, this latter interpretation has preferred for the purpose of this task, but I am open to hearing arguments in favour of the alternative.

First person statements – FP

There are three kinds of first person statements in our tagset.

First Person Commitments – **FPC** – are typically statements in which the speaker is undertaking to do something: *I will attend the meeting, I will bring the data*. Broadly speaking, if one is committing to something, the fulfilment or otherwise of said commitment can be subsequently verified (so, e.g., *I will dream of you tonight* does not count as a commitment). This includes sentences which have commitments hedged by the modal *can*, for instance: *I can finish my work before 8. I can take your place at the meeting*. Similar considerations apply to these sentences when they are embedded within another clause as the object of a verb such as *think*, even though this represents a very cautious commitment – for example, *I think I can attend the meeting on Monday.*

First Person Feelings – **FPF** – are, as the name suggests, first person statements in which the speaker's feelings and thoughts are expressed: *I am so happy to see you. I am so sorry to hear your news. I am uncertain about the agenda.* A sentence that comes up often in the data is something like *I have a question for you.* I think this should probably be included in this category, too, but am open to alternative suggestions.

First Person Other – **FPO** – should also be fairly transparent: it simply refers to all other kinds of first person statements, which mainly cover past commitments (*I sent the papers last week, I attended the meeting yesterday*) and generic statement of facts (*I am an employee of the company, I come in every morning at 9*).

Other stuff – OT and EX

All other statements that don't fall into the categories above are lumped together as **OT**, for OTher. This mainly includes third person statements that impart information (*The meeting is at 8 tomorrow. The figures for the year are ready.*) and comments directed to the hearer, so in the second person (*You must be so upset right now, You always work so hard)*. It also includes exclamations such as *Good luck on your new job!* and greetings such as *Dear Mary* or *Hi John*.

I have decided to include in this class also sentences such as *This really bothers me* or *Your decision does not make sense to me* even though these in theory express the speaker's feelings. If you feel very strongly that this is inappropriate, let me know.

EX refers to EXtraneous material – though I hope the data has been cleaned up fairly well, it's very likely that there will be bits that are irrelevant or not well-formed: filenames, urls, addresses, etc. Please tag those as **EX**.

Example 1

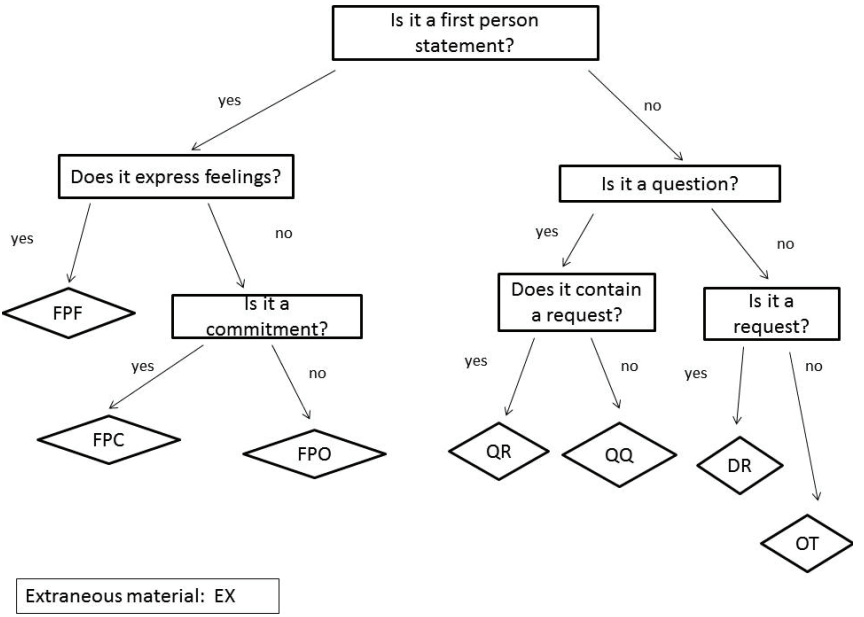
Dear Laura	OT
Thanks for your email.	OT
I will be happy to take the seminar speaker for lunch on Friday.	FPC
I'm sorry you have to miss it.	FPF
Can you tell me where the speaker is staying?	QR
Will I need to pay for his lunch?	QQ
Thanks,	OT
Rachele	OT

Example 2

Dear Davis	OT
How are you?	OT
You must be feeling stressed!	OT
Of course I can go to the meeting.	FPC
It's my pleasure.	OT
So what time should I go there?	QQ
And please tell me what to expect at the meeting.	DR
I hope you will email me back asap.	DR
Yours,	OT
Andy	OT

Example 3	
Hi Tom	OT
I know about your problem.	FPO
I think I am able to help with it.	FPC
I've dealt with this before so it's not hard for me.	FPO
I need some more information though.	FPO
So can you show me the figures?	QR
I look forward to hearing from you soon.	DR

Appendix B: Flowchart for annotation procedure



Appendix C: Tag reminders and examples

DR
Please email me back with the answers to my questions.
I am looking forward to hearing from you soon.
QR
Can you tell me where the entrance is?
Are there any tasks I should do on Monday?
QQ
Do you know who arranged your schedule?
Are you going to Nagoya?
FPF
I have a few questions to ask.
I am happy that you ask for my help.
FPC
I will be available in the morning.
I can go to the office on Monday.
FPO
I sent the papers last week.
I work at the London branch.
OT
It's my pleasure.
A map would be good for me.
You replaced me at a meeting once.
My girlfriend works there.
You seem so busy!
Good luck for your work on Monday.
EX
extraneous material