Towards a model of formal and informal address in English

Manaal Faruqui
Computer Science and Engineering
Indian Institute of Technology
Kharagpur, India
manaalfar@gmail.com

Sebastian Padó
Institute of Computational Linguistics
Heidelberg University
Heidelberg, Germany
pado@cl.uni-heidelberg.de

Abstract

Informal and formal ("T/V") address in dialogue is not distinguished overtly in modern English, e.g. by pronoun choice like in many other languages such as French ("tu"/"vous"). Our study investigates the status of the T/V distinction in English literary texts. Our main findings are: (a) human raters can label monolingual English utterances as T or V fairly well, given sufficient context; (b), a bilingual corpus can be exploited to induce a supervised classifier for T/V without human annotation. It assigns T/V at sentence level with up to 68% accuracy, relying mainly on lexical features; (c), there is a marked asymmetry between lexical features for formal speech (which are conventionalized and therefore general) and informal speech (which are text-specific).

1 Introduction

In many Indo-European languages, there are two pronouns corresponding to the English *you*. This distinction is generally referred to as the *T/V dichotomy*, from the Latin pronouns *tu* (informal, T) and *vos* (formal, V) (Brown and Gilman, 1960). The V form (such as *Sie* in German and *Vous* in French) can express neutrality or polite distance and is used to address social superiors. The T form (German *du*, French *tu*) is employed towards friends or addressees of lower social standing, and implies solidarity or lack of formality.

English used to have a T/V distinction until the 18th century, using *you* as V pronoun and *thou* for T. However, in contemporary English, *you* has taken over both uses, and the T/V distinction is not marked anymore. In NLP, this makes generation in English and translation into English easy. Conversely, many NLP tasks suffer from the lack of information about formality, e.g. the extraction of social relationships or, notably, machine translation from English into languages with a T/V distinction which involves a pronoun choice.

In this paper, we investigate the possibility to recover the T/V distinction for (monolingual) sentences of 19th and 20th-century English such as:

(1) Can I help **you**, Sir? (V)
(2) **You** are my best friend! (T)

After describing the creation of an English corpus of T/V labels via annotation projection (Section 3), we present an annotation study (Section 4) which establishes that taggers can indeed assign T/V labels to monolingual English utterances in context fairly reliably. Section 5 investigates how T/V is expressed in English texts by experimenting with different types of features, including words, semantic classes, and expressions based on Politeness Theory. We find word features to be most reliable, obtaining an accuracy of close to 70%.

2 Related Work

There is a large body of work on the T/V distinction in (socio-)linguistics and translation studies, covering in particular the conditions governing T/V usage in different languages (Kretzenbacher et al., 2006; Schüpbach et al., 2006) and the difficulties in translation (Ardila, 2003; Künzli, 2010). However, many observations from this literature are difficult to operationalize. Brown and Levinson (1987) propose a general theory of politeness which makes many detailed predictions. They assume that the pragmatic goal of being polite gives rise to general communication strategies, such as avoiding to lose face (cf. Section 5.2).

In computational linguistics, it is a common observation that for almost every language pair, there are distinctions that are expressed overtly

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 623–633,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

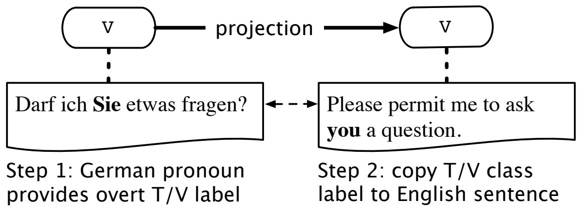
V
projection
V
Darf ich **Sie** etwas fragen?
Please permit me to ask **you** a question.
Step 1: German pronoun provides overt T/V label
Step 2: copy T/V class label to English sentence

Figure 1: T/V label induction for English sentences in a parallel corpus with annotation projection

in one language, but remain covert in the other. Examples include morphology (Fraser, 2009) and tense (Schiehlen, 1998). A technique that is often applied in such cases is *annotation projection*, the use of parallel corpora to copy information from a language where it is overtly realized to one where it is not (Yarowsky and Ngai, 2001; Hwa et al., 2005; Bentivogli and Pianta, 2005).

The phenomenon of formal and informal address has been considered in the contexts of translation into (Hobbs and Kameyama, 1990; Kanayama, 2003) and generation in Japanese (Bateman, 1988). Li and Yarowsky (2008) learn pairs of formal and informal constructions in Chinese with a paraphrase mining strategy. Other relevant recent studies consider the extraction of social networks from corpora (Elson et al., 2010). A related study is (Bramsen et al., 2011) which considers another sociolinguistic distinction, classifying utterances as “upspeak” and “downspeak” based on the social relationship between speaker and addressee.

This paper extends a previous pilot study (Faruqui and Padó, 2011). It presents more annotation, investigates a larger and better motivated feature set, and discusses the findings in detail.

3 A Parallel Corpus of Literary Texts

This section discusses the construction of T/V gold standard labels for English sentences. We obtain these labels from a parallel English–German corpus using the technique of annotation projection (Yarowsky and Ngai, 2001) sketched in Figure 1: We first identify the T/V status of German pronouns, then copy this T/V information onto the corresponding English sentence.

3.1 Data Selection and Preparation

Annotation projection requires a parallel corpus. We found commonly used parallel corpora like EUROPARL (Koehn, 2005) or the JRC Acquis corpus (Steinberger et al., 2006) to be unsuitable for our study since they either contain almost no direct address at all or, if they do, just formal address (V). Fortunately, for many literary texts from the 19th and early 20th century, copyright has expired, and they are freely available in several languages.

We identified 110 stories and novels among the texts provided by Project Gutenberg (English) and Project Gutenberg-DE (German)[1] that were available in both languages, with a total of 0.5M sentences per language. Examples are Dickens' *David Copperfield* or Tolstoy's *Anna Karenina*. We excluded plays and poems, as well as 19th-century adventure novels by Sir Walter Scott and James F. Cooper which use anachronistic English for stylistic reasons, including words that previously (until the 16th century) indicated T (“thee”, “didst”).

We cleaned the English and German novels manually by deleting the tables of contents, prologues, epilogues, as well as chapter numbers and titles occurring at the beginning of each chapter to obtain properly parallel texts. The files were then formatted to contain one sentence per line using the sentence splitter and tokenizer provided with EUROPARL (Koehn, 2005). Blank lines were inserted to preserve paragraph boundaries. All novels were lemmatized and POS-tagged using TreeTagger (Schmid, 1994).[2] Finally, they were sentence-aligned using Gargantuan (Braune and Fraser, 2010), an aligner that supports one-to-many alignments, and word-aligned in both directions using Giza++ (Och and Ney, 2003).

3.2 T/V Gold Labels for English Utterances

As Figure 1 shows, the automatic construction of T/V labels for English involves two steps.

Step 1: Labeling German Pronouns as T/V. German has three relevant personal pronouns for the T/V distinction: *du* (T), *sie* (V), and *ihr* (T/V). However, various ambiguities makes their interpretation non-straightforward.

The pronoun *ihr* can both be used for plural T address or for a somewhat archaic singular or plural V address. In principle, these usages should be distinguished by capitalization (V pronouns are generally capitalized in German), but many T instances in our corpora informal use are nevertheless capitalized. Additional, *ihr* can be the

[1] http://www.gutenberg.org, http://gutenberg.spiegel.de/

[2] It must be expected that the tagger degrades on this dataset; however we did not quantify this effect.

dative form of the 3rd person feminine pronoun *sie* (*she/her*). These instances are neutral with respect to T/V but were misanalysed by TreeTagger as instances of the T/V lemma *ihr*. Since TreeTagger does not provide person information, and we did not want to use a full parser, we decided to omit *ihr/Ihr* from consideration.[3]

Of the two remaining pronouns (*du* and *sie*), *du* expresses (singular) T. A minor problem is presented by novels set in France, where *du* is used as an nobiliary particle. These instances can be recognised reliably since the names before and after *du* are generally unknown to the German tagger. Thus we do not interpret *du* as T if the word preceding or succeeding it has "unknown" as its lemma.

The V pronoun, *sie*, doubles as the pronoun for third person (*she/they*) when not capitalized. We therefore interpret only capitalized instances of *Sie* as V. Furthermore, we ignore utterance-initial positions, where all words are capitalized. This is defined as tokens directly after a sentence boundary (POS `$.`) or after a bracket (POS `$(`).

These rules concentrate on precision rather than recall. They leave many instances of German second person pronouns unlabeled; however, this is not a problem since we do not currently aim at obtaining complete coverage on the English side of our parallel corpus. From the 0.5M German sentences, about 14% of the sentences were labeled as T or V (37K for V and 28K for T). In a random sample of roughly 300 German sentences which we analysed, we did not find any errors. This puts the precision of our heuristics at above 99%.

Step 2: Annotation Projection. We now copy the information over onto the English side. We originally intended to transfer T/V labels between German and English word-aligned pronouns. However, we pronouns are not necessarily translated into pronouns; additionally, we found word alignment accuracy for pronouns to be far from perfect, due to the variability in function word translation. For these reason, we decided to look at T/V labels at the level of complete sentences, ignoring word alignment. This is generally unproblematic – address is almost always consistent within sentences: of the 65K German sentences with T or V labels, only 269 ($< 0.5\%$) contain both T and V. Our projection on the English side results in 25K V and 18K T sentences[4], of which 255 (0.6%) are labeled as both T and V. We exclude these sentences.

Comparison	No context	In context
A1 vs. A2	75% (.49)	79% (.58)
A1 vs. GS	60% (.20)	70% (.40)
A2 vs. GS	65% (.30)	76% (.52)
(A1 ∩ A2) vs. GS	67% (.34)	79% (.58)

Table 1: Manual annotation for T/V on a 200-sentence sample. Comparison among human annotators (A1 and A2) and to projected gold standard (GS). All cells show raw agreement and Cohen's κ (in parentheses).

Note that this strategy relies on the *direct correspondence assumption* (Hwa et al., 2005), that is, it assumes that the T/V status of an utterance is not changed in translation. We believe that this is a reasonable assumption, given that T/V is determined by the social relation between interlocutors; but see Section 4 for discussion.

3.3 Data Splitting

Finally, we divided our English data into training, development and test sets with 74 novels (26K sentences), 19 novels (9K sentences) and 13 novels (8K sentences), respectively. The corpus is available for download at `http://www.nlpado.de/~sebastian/data.shtml`.

4 Human Annotation of T/V for English

This section investigates how well the T/V distinction can be made in English by human raters, and on the basis of what information. Two annotators with near native-speaker competence in English were asked to label 200 random sentences from the training set as T or V. Sentences were first presented in isolation ("no context"). Subsequently, they were presented with three sentences pre- and post-context each ("in context").

Table 1 shows the results of the annotation study. The first line compares the annotations of the two annotators against each other (inter-annotator agreement). The next two lines compare the taggers' annotations against the gold standard labels projected from German (GS). The last line compares the annotator-assigned labels to the GS for the instances on which the annotators agree. For all cases, we report raw accuracy and Cohen's κ (1960), i.e. chance-corrected agreement.

[3] Instances of *ihr* as possessive pronoun occurred as well, but could be filtered out on the basis of the POS tag.

[4] Our sentence aligner supports one-to-many alignments and often aligns single German to multiple English sentences.

We first observe that the T/V distinction is considerably more difficult to make for individual sentences (no context) than when the discourse is available. In context, inter-annotator agreement increases from 75% to 79%, and agreement with the gold standard rises by 10%. It is notable that the two annotators agree worse with one another than with the gold standard (see below for discussion). On those instances where they agree, Cohen's κ reaches 0.58 in context, which is interpreted as approaching good agreement (Fleiss, 1981). Although far from perfect, this inter-annotator agreement is comparable to results for the annotation of fine-grained word sense or sentiment (Navigli, 2009; Bermingham and Smeaton, 2009).

An analysis of disagreements showed that many sentences can be uttered in both T and V contexts and cannot be labeled without context:

(3) "And perhaps sometime **you** may see her."

This case (gold label: V) is disambiguated by the previous sentence which indicates a hierarchical social relation between speaker and addressee:

(4) "And she is a sort of relation of **your lordship's**," said Dawson. ...

Still, even a three-sentence window is often not sufficient, since the surrounding sentences may be just as uninformative. In these cases, more global information about the situation is necessary. Even with perfect information, however, judgments can sometimes deviate, as there are considerable "grey areas" in T/V usage (Kretzenbacher et al., 2006).

In addition, social rules like T/V usage vary in time and between countries (Schüpbach et al., 2006). This helps to explain why annotators agree better with one another than with the gold standard: 21st century annotators tend to be unfamiliar with 19th century T/V usage. Consider this example from a book written in second person perspective:

(5) Finally, **you** acquaint Caroline with the fatal result: she begins by consoling **you**. "One hundred thousand francs lost! We shall have to practice the strictest economy", **you** imprudently add.[5]

Here, the author and translator use V to refer to the reader, while today's usage would almost certainly be T, as presumed by both annotators. Conversations between lovers or family members form another example, where T is modern usage, but the novels tend to use V:

(6) [...] she covered her face with the other to conceal her tears. "Corinne!", said Oswald, "Dear Corinne! My absence has then rendered **you** unhappy!"[6]

In sum, our annotation study establishes that the T/V distinction, although not realized by different pronouns in English, can be recovered manually from text, provided that discourse context is available. A substantial part of the errors is due to social changes in T/V usage.

5 Monolingual T/V Modeling

The second part of the paper explores the automatic prediction of the T/V distinction for English sentences. Given the ability to create an English training corpus with T/V labels with the annotation projection methods described in Section 3.2, we can phrase T/V prediction for English as a standard supervised learning task. Our experiments have a twin motivation: (a), on the NLP side, we are mainly interested in obtaining a robust classifier to assign the labels T and V to English sentences; (b), on the sociolinguistic side, we are interested in investigating through which features the categories T and V are expressed in English.

5.1 Classification Framework

We phrase T/V labeling as a binary classification task at the sentence level, performing the classification with L2-regularized logistic regression using the LibLINEAR library (Fan et al., 2008). Logistic regression defines the probability that a binary response variable y takes some value as a logit-transformed linear combination of the features f_i, each of which is assigned a coefficient β_i.

$$p(y=1) = \frac{1}{1+e^{-z}} \text{ with } z = \sum_i \beta_i f_i \qquad (7)$$

Regularization incorporates the size of the coefficient vector β into the objective function, subtracting it from the likelihood of the data given the model. This allows the user to trade faithfulness to the data against generalization.[7]

[5] H. de Balzac: Petty Troubles of Married Life

[6] A.L.G. de Staël: Corinne

[7] We use LIBLINEAR's default parameters and set the cost (regularization) parameter to 0.01.

$\frac{p(C\|V)}{p(C\|T)}$	Words
4.59	Mister, sir, Monsieur, sirrah, ...
2.36	Mlle., Mr., M., Herr, Dr., ...
1.60	Gentlemen, patients, rascals, ...

Table 2: 3 of the 400 clustering-based semantic classes (classes most indicative for V)

5.2 Feature Types

We experiment with three features types that are candidates to express the T/V English distinction.

Word Features. The intuition to use word features draws on the parallel between T/V and information retrieval tasks like document classification: some words are presumably correlated with formal address (like titles), while others should indicate informal address (like first names). In a preliminary experiment, we noticed that in the absence of further constraints, many of the most indicative features are names of persons from particular novels which are systematically addressed formally (like Phileas Fogg from J. Vernes' *Around the world in eighty days*) or informally (like Mowgli, Baloo, and Bagheera from R. Kipling's *Jungle Book*). These features clearly do not generalize to new books. We therefore added a constraint to remove all features which did not occur in at least three novels. To reduce the number of word features to a reasonable order of magnitude, we also performed a χ^2-based feature selection (Manning et al., 2008) on the training set. Preliminary experiments established that selecting the top 800 word features yielded a model with good generalization.

Semantic Class Features. Our second feature type is semantic class features. These can be seen as another strategy to counteract the sparseness at the level of word features. We cluster words into 400 semantic classes on the basis of distributional and morphological similarity features which are extracted from an unlabeled English collection of Gutenberg novels comprising more than 100M tokens, using the approach by Clark (2003). These features measure how similar tokens are to one another in terms of their occurrences in the document and are useful in Named Entity Recognition (Finkel and Manning, 2009). As features in the T/V classification of a given sentence, we simply count for each class the number of tokens in this class present in the current sentence. For illustration, Table 2 shows the three classes most indicative for V, ranked by the ratio of probabilities for T and V, estimated on the training set.

Politeness Theory Features. The third feature type is based on the Politeness Theory (Brown and Levinson, 1987). Brown and Levinson's prediction is that politeness levels will be detectable in concrete utterances in a number of ways, e.g. a higher use of conjunctive or hedges in polite speech. Formal address (i.e., V as opposed to T) is one such expression. Politeness Theory therefore predicts that other politeness indicators should correlate with the T/V classification. This holds in particular for English, where pronoun choice is unavailable to indicate politeness.

We constructed 16 features on the basis of Politeness Theory predictions, that is, classes of expressions indicating either formality or informality. From a computational perspective, the problem with Politeness Theory predictions is that they are only described qualitatively and by example, without detailed lists. For each feature, we manually identified around 10 words or multi-word relevant expressions. Table 3 shows these 16 features with their intended classes and some example expressions. Similar to the semantic class features, the value of each politeness feature is the sum of the frequencies of its members in a sentence.

5.3 Context: Size and Type

As our annotation study in Section 4 found, context is crucial for human annotators, and this presumably carries over to automatic methods human annotators: if the features for a sentence are computed just on that sentence, we will face extremely sparse data. We experiment with symmetrical window contexts, varying the size between $n = 0$ (just the target sentence) and $n = 10$ (target sentence plus 10 preceding and 10 succeeding sentences).

This kind of simple "sentence context" makes an important oversimplification, however. It lumps together material from different speech turns as well as from "narrative" sentences, which may generate misleading features. For example, narrative sentences may refer to protagonists by their full names including titles (strong features for V) even when these protagonists are in T-style conversations:

(8) "You are the love of my life", said Sir Phileas Fogg.[8] (T)

[8] J. Verne: Around the world in 80 days

Class	Example expressions	Class	Example expressions
Inclusion (T)	let's, shall we	Exclamations (T)	hey, yeah
Subjunctive I (T)	can, will	Subjunctive II (V)	could, would
Proximity (T)	this, here	Distance (V)	that, there
Negated question (V)	didn't I, hasn't it	Indirect question (V)	would there, is there
Indefinites (V)	someone, something	Apologizing (V)	bother, pardon
Polite adverbs (V)	marvellous, superb	Optimism (V)	I hope, would you
Why + modal (V)	why would(n't)	Impersonals (V)	necessary, have to
Polite markers (V)	please, sorry	Hedges (V)	in fact, I guess

Table 3: 16 Politeness theory-based features with intended classes and example expressions

Example (8) also demonstrates that narrative material and direct speech may even be mixed within individual sentences.

For these reasons, we introduce an alternative concept of context, namely *direct speech context*, whose purpose is to exclude narrative material. We compute direct speech context in two steps: (a), segmentation of sentences into chunks that are either completely narrative or speech, and (b), labeling of chunks with a classifier that distinguishes these two classes. The segmentation step (a) takes place with a regular expression that subdivides sentences on every occurrence of quotes (“ , ” , ’ , ‘, etc.). As training data for the classification step (b), we manually tagged 1000 chunks from our training data as either B-DS (begin direct speech), I-DS (inside direct speech) and O (outside direct speech, i.e. narrative material).[9] We used this dataset to train the CRF-based sequence tagger Mallet (McCallum, 2002) using all tokens, including punctuation, as features.[10] This tagger is used to classify all chunks in our dataset, resulting in output like the following example:

(9)
(B-DS) “I am going to see his Ghost!
(I-DS) It will be his Ghost not him!”
(O) Mr. Lorry quietly chafed the hands that held his arm.[11]

Direct speech chunks belonging to the same sentence are subsequently recombined.

We define the direct speech context of size n for a given sentence as the n preceding and following direct speech chunks that are labeled B-DS or I-DS while skipping any chunks labeled O. Note that this definition of direct speech context still lumps together utterances by different speakers and can therefore yield misleading features in the case of asymmetric conversational situations, in addition to possible direct speech misclassifications.

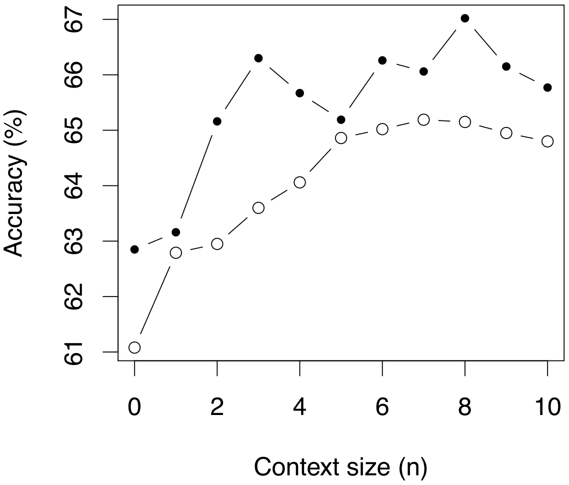

Figure 2: Accuracy vs. number of sentences in context (empty circles: sentence context; solid circles: direct speech context)

6 Experimental Evaluation

6.1 Evaluation on the Development Set

We first perform model selection on the development set and then validate our results on the test set (cf. Section 3.3).

Influence of Context. Figure 2 shows the influence of size and type of context, using only words as features. Without context, we obtain a performance of 61.1% (sentence context) and of 62.9% (direct speech context). These numbers beat the random baseline (50.0%) and the frequency baseline (59.1%). The addition of more context further improves performance substantially for both context types. The ideal context size is fairly large, namely 7 sentences and 8 direct speech chunks, re-

[9] The labels are chosen after IOB notation conventions (Ramshaw and Marcus, 1995).

[10] We also experimented with rule-based chunk labeling based on quotes, but found the use of quotes too inconsistent.

[11] C. Dickens: A tale of two cities.

Model	Accuracy
Random Baseline	50.0
Frequency Baseline	59.1
Words	**67.0****
SemClass	57.5
PoliteClass	59.6
Words + SemClass	66.6**
Words + PoliteClass	66.4**
Words + PoliteClass + SemClass	66.2**
Raw human IAA (no context)	75.0
Raw human IAA (in context)	79.0

Table 4: T/V classification accuracy on the development set (direct speech context, size 8). **: Significant difference to frequency baseline (p<0.01)

spectively. This indicates that sparseness is indeed a major challenge, and context can become large before the effects mentioned in Section 5.3 counteract the positive effect of more data. Direct speech context outperforms sentence context throughout, with a maximum accuracy of 67.0% as compared to 65.2%, even though it shows higher variation, which we attribute to the less stable nature of the direct speech chunks and their automatically created labels. From now on, we adopt a direct speech context of size 8 unless specified differently.

Influence of Features. Table 4 shows the results for different feature types. The best model (word features only) is highly significantly better than the frequency baseline (which it beats by 8%) as determined by a bootstrap resampling test (Noreen, 1989). It gains 17% over the random baseline, but is still more than 10% below inter-annotator agreement in context, which is often seen as an upper bound for automatic models.

Disappointingly, the comparison of the feature groups yields a null result: We are not able to improve over the results for just word features with either the semantic class or the politeness features. Neither feature type outperforms the frequency baseline significantly (p>0.05). Combinations of the different feature types also do worse than just words. The differences between the best model (just words) and the combination models are all not significant (p>0.05). These negative results warrant further analysis. It follows in Section 6.3.

6.2 Results on the Test Set

Table 5 shows the results of evaluating models with the best feature set and with different context sizes on the test set, in order to verify that we did not overfit on the development set when picking the best model. The tendencies correspond well to the development set: the frequency baseline is almost identical, as are the results for the different models. The differences to the development set are all equal to or smaller than 1% accuracy, and the best result at 67.5% is 0.5% better than on the development set. This is a reassuring result, as our model appears to generalize well to unseen data.

Model	Accuracy	Δ to dev set
Frequency baseline	59.3	+ 0.2
Words (no context)	62.5	- 0.4
Words (context size 6)	67.3	+ 1.0
Words (context size 8)	**67.5**	+ 0.5
Words (context size 10)	66.8	+ 1.0

Table 5: T/V classification accuracy on the test set and differences to dev set results (direct speech context)

6.3 Analysis by Feature Types

The results from Section 6.1 motivate further analysis of the individual feature types.

Analysis of Word Features. Word features are by far the most effective features. Table 6 lists the top twenty words indicating T and V (ranked by the ratio of probabilities for the two classes on the training set). The list still includes some proper names like *Vrazumihin* or *Louis-Gaston* (even though all features have to occur in at least three novels), but they are relatively infrequent. The most prominent indicators for the formal class V are titles (*monsieur*, *(ma)'am*) and instances of formulaic language (*Permit (me)*, *Excuse (me)*). There are also some terms which are not straightforward indicators of formal address (*angelic*, *stubbornness*), but are associated with a high register.

There is a notable asymmetry between T and V. The word features for T are considerably more difficult to interpret. We find some forms of earlier period English (*thee, hast, thou, wilt*) that result from occasional archaic passages in the novels as well first names (*Louis-Gaston*, *Justine*). Nevertheless, most features are not straightforward to connect to specifically informal speech.

Analysis of Semantic Class Features. We ranked the semantic classes we obtained by distributional clustering in a similar manner to the word features. Table 2 shows the top three classes indicative for V. Almost all others of the 400 clusters do not have a strong formal/informal association

Top 20 words for V		Top 20 words for T	
Word w	$\frac{P(w\|V)}{P(w\|T)}$	Word w	$\frac{P(w\|T)}{P(w\|V)}$
Excuse	36.5	thee	94.3
Permit	35.0	amenable	94.3
'ai	29.2	stuttering	94.3
'am	29.2	guardian	94.3
stubbornness	29.2	hast	92.0
flights	29.2	Louis-Gaston	92.0
monsieur	28.6	lease-making	92.0
Vrazumihin	28.6	melancholic	92.0
mademoiselle	26.5	ferry-boat	92.0
angelic	26.5	Justine	92.0
Allow	24.5	Thou	66.0
madame	21.2	responsibility	63.8
delicacies	21.2	thou	63.8
entrapped	21.2	Iddibal	63.8
lack-a-day	21.2	twenty-fifth	63.8
ma	21.0	Chic	63.8
duke	18.0	allegiance	63.8
policeman	18.0	Jouy	63.8
free-will	18.0	wilt	47.0
Canon	18.0	shall	47.0

Table 6: Most indicative word features for T or V

but mix formal, informal, and neutral vocabulary. This tendency is already apparent in class 3: *Gentlemen* is clearly formal, while *rascals* is informal. *patients* can belong to either class. Even in class 1, we find *Sirrah*, a contemptuous term used in addressing a man or boy with a low formality score ($p(w|V)/p(w|T) = 0.22$). From cluster 4 onward, none of the clusters is strongly associated with either V or T ($p(c|V)/p(c|T) \approx 1$).

Our interpretation of these observations is that in contrast to text categorization, there is no clear-cut topical or domain difference between T and V: both categories co-occur with words from almost any domain. In consequence, semantic classes do not, in general, represent strong unambiguous indicators. Similar to the word features, the situation is worse for T than for V: there still are reasonably strong features for V, the "marked" case, but it is more difficult to find indicators for T.

Analysis of politeness features. A major reason for the ineffectiveness of the Politeness Theory-based features seems to be their low frequency: in the best model, with a direct speech context of size 8, only an average of 7 politeness features was active for any given sentence. However, frequency was not the only problem – the politeness features were generally unable to discriminate well between T and V. For all features, the values of $p(f|V)/p(f|T)$ are between 0.9 and 1.3, that is, the features were only weakly indicative of one of the classes. Furthermore, not all features turned out to be indicative of the class we designed them for. The best indicator for V was the Indefinites feature (*somehow, someone* cf. Table 3), as expected. In contrast, the best indicator for T was the Negation question feature which was supposedly an indicator for V (*didn't I, haven't we*).

A majority of politeness features (13 of the 16) had $p(f|V)/p(f|T)$ values above 1, that is, were indicative for the class V. Thus for this feature type, like for the others, it appears to be more difficult to identify T than to identify V. This negative result can be attributed at least in part to our method of hand-crafting lists of expressions for these features. The inadvertent inclusion of overly general terms V might be responsible for the features' inability to discriminate well, while we have presumably missed specific terms which has hurt coverage. This situation may in the future be remedied with the semi-automatic acquisition of instantiations of politeness features.

6.4 Analysis of Individual Novels

One possible hypothesis regarding the difficulty of finding indicators for the class T is that indicators for T tend to be more novel-specific than indicators for V, since formal language is more conventionalized (Brown and Levinson, 1987). If this were the case, then our strategy of building well-generalizing models by combining text from different novels would naturally result in models that have problems with picking up T features.

To investigate this hypothesis, we trained models with the best parameters as before (8-sentence direct speech context, words as features). However, this time we trained novel-specific models, splitting each novel into 50% training data and 50% testing data. We required novels to contain more than 200 labeled sentences. This ruled out most short stories, leaving us with 7 novels in the test set. The results are shown in Table 7 and show a clear improvement. The accuracy is 13% higher than in our main experiment (67% vs. 80%), even though the models were trained on considerably less data. Six of the seven novels perform above the 67.5% result from the main experiment.

The top-ranked features for T and V show a much higher percentage of names for both T and V than in the main experiment. This is to be ex-

Novel	Accuracy
H. Beecher-Stove: Uncle Tom's Cabin	90.0
J. Spyri: Cornelli	88.3
E. Zola: Lourdes	83.9
H. de Balzac: Cousin Pons	82.3
C. Dickens: The Pickwick Papers	77.7
C. Dickens: Nicholas Nickleby	74.8
F. Hodgson Burnett: Little Lord	61.6
All (micro average)	80.0

Table 7: T/V prediction models for individual novels (50% of each novel for training and 50% testing)

pected, since this experiment does not restrict itself to features that occurred in at least three novels. The price we pay for this is worse generalization to other novels. There is also still a T/V asymmetry: more top features are shared among the V lists of individual novels and with the main experiment V list than on the T side. Like in the main experiment (cf. Section 6.3), V features indicate titles and other features of elevated speech, while T features mostly refer to novel-specific protagonists and events. In sum, these results provide evidence for a difference in status of T and V.

7 Discussion and Conclusions

In this paper, we have studied the distinction between formal and information (T/V) address, which is not expressed overtly through pronoun choice or morphosyntactic marking in modern English. Our hypothesis was that the T/V distinction can be recovered in English nevertheless. Our manual annotation study has shown that annotators can in fact tag monolingual English sentences as T or V with reasonable accuracy, but only if they have sufficient context. We exploited the overt information from German pronouns to induce T/V labels for English and used this labeled corpus to train a monolingual T/V classifier for English. We experimented with features based on words, semantic classes, and Politeness Theory predictions.

With regard to our NLP goal of building a T/V classifier, we conclude that T/V classification is a phenomenon that can be modelled on the basis of corpus features. A major factor in classification performance is the inclusion of a wide context to counteract sparse data, and more sophisticated context definitions improve results. We currently achieve top accuracies of 67%-68%, which still leave room for improvement. We next plan to couple our T/V classifier with a machine translation system for a task-based evaluation on the translation of direct address into German and other languages with different T/V pronouns.

Considering our sociolinguistic goal of determining the ways in which English realizes the T/V distinction, we first obtained a negative result: only word features perform well, while semantic classes and politeness features do hardly better than a frequency baseline. Notably, there are no clear "topical" divisions between T and V, like for example in text categorization: almost all words are very weakly correlated with either class, and semantically similar words can co-occur with different classes. Consequently, distributionally determined semantic classes are not helpful for the distinction. Politeness features are difficult to operationalize with sufficiently high precision and recall.

An interesting result is the asymmetry between the linguistic features for V and T at the lexical level. V language appears to be more conventionalized; the models therefore identified formulaic expressions and titles as indicators for V. On the other hand, very few such generic features exist for the class T; consequently, the classifier has a hard time learning good discriminating and yet generic features. Those features that are indicative of T, such as first names, are highly novel-specific and were deliberately excluded from the main experiment. When we switched to individual novels, the models picked up such features, and accuracy increased – at the cost of lower generalizability between novels. A more technical solution to this problem would be the training of a single-class classifier for V, treating T as the "default" class (Tax and Duin, 1999).

Finally, an error analysis showed that many errors arise from sentences that are too short or unspecific to determine T or V reliably. This points to the fact that T/V should not be modelled as a sentence-level classification task in the first place: T/V is not a choice made for each sentence, but one that is determined once for each pair of interlocutors and rarely changed. In future work, we will attempt to learn social networks from novels (Elson et al., 2010), which should provide constraints on all instances of communication between a speaker and an addressee. However, the big – and unsolved, as far as we know – challenge is to automatically assign turns to interlocutors, given the varied and often inconsistent presentation of direct speech turns in novels.

References

John Ardila. 2003. (Non-Deictic, Socio-Expressive) T-/V-Pronoun Distinction in Spanish/English Formal Locutionary Acts. *Forum for Modern Language Studies*, 39(1):74–86.

John A. Bateman. 1988. Aspects of clause politeness in Japanese: An extended inquiry semantics treatment. In *Proceedings of ACL*, pages 147–154, Buffalo, New York.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Journal of Natural Language Engineering*, 11(3):247–261.

Adam Bermingham and Alan F. Smeaton. 2009. A study of inter-annotator agreement for opinion retrieval. In *Proceedings of ACM SIGIR*, pages 784–785.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of ACL/HLT*, pages 773–782, Portland, OR.

Fabienne Braune and Alexander Fraser. 2010. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Coling 2010: Posters*, pages 81–89, Beijing, China.

Roger Brown and Albert Gilman. 1960. The pronouns of power and solidarity. In Thomas A. Sebeok, editor, *Style in Language*, pages 253–277. MIT Press, Cambridge, MA.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Number 4 in Studies in Interactional Sociolinguistics. Cambridge University Press.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of EACL*, pages 59–66, Budapest, Hungary.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of ACL*, pages 138–147, Uppsala, Sweden.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Manaal Faruqui and Sebastian Padó. 2011. "I Thou Thee, Thou Traitor": Predicting formal vs. informal address in English literature. In *Proceedings of ACL/HLT 2011*, pages 467–472, Portland, OR.

Jenny Rose Finkel and Christopher D. Manning. 2009. Nested named entity recognition. In *Proceedings of EMNLP*, pages 141–150, Singapore.

Joseph L. Fleiss. 1981. *Statistical methods for rates and proportions*. John Wiley, New York, 2nd edition.

Alexander Fraser. 2009. Experiments in morphosyntactic processing for translating to and from German. In *Proceedings of the EACL MT workshop*, pages 115–119, Athens, Greece.

Jerry Hobbs and Megumi Kameyama. 1990. Translation by abduction. In *Proceedings of COLING*, pages 155–161, Helsinki, Finland.

Rebecca Hwa, Philipp Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Journal of Natural Language Engineering*, 11(3):311–325.

Hiroshi Kanayama. 2003. Paraphrasing rules for automatic evaluation of translation into Japanese. In *Proceedings of the Second International Workshop on Paraphrasing*, pages 88–93, Sapporo, Japan.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Heinz L. Kretzenbacher, Michael Clyne, and Doris Schüpbach. 2006. Pronominal Address in German: Rules, Anarchy and Embarrassment Potential. *Australian Review of Applied Linguistics*, 39(2):17.1–17.18.

Alexander Künzli. 2010. Address pronouns as a problem in French-Swedish translation and translation revision. *Babel*, 55(4):364–380.

Zhifei Li and David Yarowsky. 2008. Mining and modeling relations between formal and informal Chinese phrases from web corpora. In *Proceedings of EMNLP*, pages 1031–1040, Honolulu, Hawaii.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 1st edition.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Roberto Navigli. 2009. Word Sense Disambiguation: a survey. *ACM Computing Surveys*, 41(2):1–69.

Eric W. Noreen. 1989. *Computer-intensive Methods for Testing Hypotheses: An Introduction*. John Wiley and Sons Inc.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Proceeding of the 3rd ACL Workshop on Very Large Corpora*, Cambridge, MA.

Michael Schiehlen. 1998. Learning tense translation from bilingual corpora. In *Proceedings of ACL/COLING*, pages 1183–1187, Montreal, Canada.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Doris Schüpbach, John Hajek, Jane Warren, Michael Clyne, Heinz Kretzenbacher, and Catrin Norrby. 2006. A cross-linguistic comparison of address pronoun use in four European languages: Intralingual and interlingual dimensions. In *Proceedings of the Annual Meeting of the Australian Linguistic Society*, Brisbane, Australia.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, and Dan Tufis. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC*, pages 2142–2147, Genoa, Italy.

David M. J. Tax and Robert P. W. Duin. 1999. Support vector domain description. *Pattern Recognition Letters*, 20:1191–1199.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of NAACL*, pages 200–207, Pittsburgh, PA.