"I Thou Thee, Thou Traitor": Predicting Formal vs. Informal Address in English Literature

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

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

Abstract

In contrast to many languages (like Russian or French), modern English does not distinguish formal and informal ("T/V") address overtly, for example by pronoun choice. We describe an ongoing study which investigates to what degree the T/V distinction is recoverable in English text, and with what textual features it correlates. Our findings are: (a) human raters can label English utterances as T or V fairly well, given sufficient context; (b), lexical cues can predict T/V almost at human level.

1 Introduction

In many Indo-European languages, such as French, German, or Hindi, 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 can express neutrality or polite distance and is used to address socially superiors. The T form is employed for friends or addressees of lower social standing, and implies solidarity or lack of formality. Some examples for V pronouns in different languages are *Sie* (German), *Vous* (French), and आप [*Aap*] (Hindi). The corresponding T pronouns are *du*, *tu*, and तुम [*tum*].

English used to have a T/V distinction until the 18th century, using *you* as V and *thou* as T pronoun. However, in contemporary English, *you* has taken over both uses, and the T/V distinction is not marked morphosyntactically any more. This makes generation in English and translation into English easy. Conversely, the extraction of social information from texts, and translation from English into languages with a T/V distinction is very difficult.

In this paper, we investigate the possibility to recover the T/V distinction based on monolingual English text. We first demonstrate that annotators can assign T/V labels to English utterances fairly well (but not perfectly). To identify features that indicate T and V, we create a parallel English–German corpus of literary texts and preliminarily identify features that correlate with formal address (like titles, and formulaic language) as well as informal address. Our results could be useful, for example, for MT from English into languages that distinguish T and V, although we did not test this prediction with the limits of a short paper.

From a Natural Language Processing point of view, the recovery of T/V information is an instance of a more general issue in cross-lingual NLP and machine translation where for almost every language pair, there are distinctions that are not expressed overtly in the source language, but are in the target language, and must therefore be recovered in some way. Other examples from the literature include morphology (Fraser, 2009) and tense (Schiehlen, 1998). The particular problem of T/V address has been considered in the context of translation into Japanese (Hobbs and Kameyama, 1990; Kanayama, 2003) and generation (Bateman, 1988), but only on the context of knowledge-rich methods. As for data-driven studies, we are only aware of Li and Yarowsky's (2008) work, who learn pairs of formal and informal constructions in Chinese where T/V is expressed mainly in construction choice.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 467–472,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

Naturally, there is a large body of work on T/V in (socio-)linguistics and translation science, covering in particular the conditions governing T/V use in different languages (Kretzenbacher et al., 2006; Schüpbach et al., 2006) and on the difficulties in translating them (Ardila, 2003; Künzli, 2010). However, these studies are generally not computational in nature, and most of their observations and predictions are difficult to operationalize.

2 A Parallel Corpus of Literary Texts

2.1 Data Selection

We chose literary texts to build a parallel corpus for the investigation of the T/V distinction. The main reason is that commonly used non-literary collections like EUROPARL (Koehn, 2005) consist almost exclusively of formal interactions and are therefore of no use to us. Fortunately, many 18th and 19th century texts are freely available in several languages.

We identified 115 novels among the texts provided by Project Gutenberg (English) and Project Gutenberg-DE (German) that were available in both languages, with a total of 0.5M sentences per language.[1] Examples include Dickens' *David Copperfield* or Tolstoy's *Anna Karenina*. We decided to exclude plays and poems as they often include partial sentences and structures that are difficult to align.

2.2 Data Preparation

As the German and English novels come from two different websites, they were not coherent in their structure. They were first manually cleaned by deleting the index, prologue, epilogue and Gutenberg license from the beginning and end of the files. To some extent the chapter numbers and titles occurring at the beginning of each chapter were cleared as well. The files were then formatted to contain one sentence per line and a blank line was inserted to preserve the segmentation information.

The sentence splitter and tokenizer provided with EUROPARL (Koehn, 2005) were used. We obtained a comparable corpus of English and German novels using the above pre-processing. The files in the corpus were sentence-aligned using Gargantuan (Braune and Fraser, 2010), an aligner that supports one-to-many alignments. After obtaining the sentence aligned corpus we computed word alignments in both English to German and German to English directions using Giza++ (Och and Ney, 2003). The corpus was lemmatized and POS-tagged using TreeTagger (Schmid, 1994). We did not apply a full parser to keep processing as efficient as possible.

ID	Position	Lemma	Cap	Category
(1)	any	du	any	T
(2)	non-initial	sie	yes	V
(3)	non-initial	ihr	no	T
(4)	non-initial	ihr	yes	V

Table 1: Rules for T/V determination for German personal pronouns. (Cap: Capitalized)

2.3 T/V Gold Labels for English Utterances

The goal of creating our corpus is to enable the investigation of contextual correlates of T/V in English. In order to do this, we need to decide for as many English utterances in our corpus as possible whether they instantiate formal or informal address. Given that we have a parallel corpus where the German side overtly realizes T and V, this is a classical case of annotation projection (Yarowsky and Ngai, 2001): We transfer the German T/V information onto the English side to create an annotated English corpus. This allows us to train and evaluate a monolingual English classifier for this phenomenon. However, two problems arise on the way:

Identification of T/V in German pronouns. German has three relevant personal pronouns: *du*, *sie*, and *ihr*. These pronouns indicate T and V, but due to their ambiguity, it is impossible to simply interpret their presence or absense as T or V. We developed four simple disambiguation rules based on position on the sentence and capitalization, shown in Table 1.

The only unambiguous pronoun is *du*, which expresses (singular) T (Rule 1). The V pronoun for singular, *sie*, doubles as the pronoun for third person (singular and plural), which is neutral with respect to T/V. Since TreeTagger does not provide person information, the only indicator that is available is capitalization: *Sie* is 2nd person V. However, since all words are capitalized in utterance-initial positions, we only assign the label V in non-initial positions

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

(Rule 2).[2]

Finally, *ihr* is also ambiguous: non-capitalized, it is used as T plural (Rule 3); capitalized, it is used as an archaic alternative to *Sie* for V plural (Rule 4).

These rules leave a substantial number of instances of German second person pronouns unlabeled; we cover somewhat more than half of all pronouns. In absolute numbers, from 0.5M German sentences we obtained about 15% labeled sentences (45K for V and 30K for T). However, this is not a fundamental problem, since we subsequently used the English data to train a classifier that is able to process any English sentence.

Choice of English units to label. On the German side, we assign the T/V labels to pronouns, and the most straightforward way of setting up annotation projection would be to label their word-aligned English pronouns as T/V. However, pronouns are not necessarily translated into pronouns; additionally, we found word alignment accuracy for pronouns, as a function of word class, to be far from perfect. For these reasons, we decided to treat *complete sentences* as either T or V. This means that sentence alignment is sufficient for projection, but English sentences can receive conflicting labels, if a German sentence contains both a T and a V label. However, this occurs very rarely: of the 76K German sentences with T or V pronouns, only 515, or less than 1%, contain both. Our projection on the English side results in 53K V and 35K T sentences, of which 731 are labeled as both T and V.[3]
Finally, from the English labeled sentences we extracted a training set with 72 novels (63K sentences) and a test set with 21 novels (15K sentences).[4]

3 Experiment 1: Human Annotation

The purpose of our first experiment is to investigate how well the T/V distinction can be made in English by human raters, and on the basis of what information. We extracted 100 random sentences from the training set. Two annotators with advanced knowledge of English were asked to label these sentences as T or V. In a first round, the sentences were presented in isolation. In a second round, the sentences were presented with three sentences pre-context and three sentences post-context. The results in Table 2 show that it is fairly difficult to annotate the T/V distinction on individual sentences since it is not expressed systematically. At the level of small discourses, the distinction can be made much more confidently: In context, average agreement with the gold standard rises from 64% to 70%, and raw inter-annotator agreement goes up from 68% to 81%.

	Acc (Ann1)	Acc (Ann2)	IAA
No context	63	65	68
In context	70	69	81

Table 2: Manual annotation for T/V on a 100-sentence sample (Acc: Accuracy, IAA: Inter-annotator agreement)

Concerning the interpretation of these findings, we note that the two taggers were both native speakers of languages which make an overt T/V distinction. Thus, our present findings cannot be construed as firm evidence that English speakers make a distinction, even if implicitly. However, they demonstrate at least that native speakers of such languages can recover the distinction based solely on the clues in English text.

An analysis of the annotation errors showed that many individual sentences can be uttered in both T and V situations, making it impossible to label them in isolation:

(1) "And perhaps sometime you may see her."

This case (gold label: V) is however disambiguated by looking at the previous sentence, which indicates the social relation between speaker and addressee:

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

Still, a three-sentence window is often not sufficient, since the surrounding sentences may be just as uninformative. In these cases, global information about the situation would be necessary.

A second problem is the age of the texts. They are often difficult to label because they talk about social situations that are unfamiliar to modern speakers (as

[2] An initial position is defined as a position after a sentence boundary (POS "$.") or after a bracket (POS "$(").

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

[4] The corpus can be downloaded for research purposes from http://www.nlpado.de/~sebastian/data.shtml.

between aristocratic friends) or where the usage has changed (as in married couples).

4 Experiment 2: Statistical Modeling

Task Setup. In this pilot modeling experiment, we explore a (limited) set of cues which can be used to predict the V vs. T dichotomy for English sentences. Specifically, we use local words (i.e. information present within the current sentence – similar to the information available to the human annotators in the "No context" condition of Experiment 1). We approach the task by supervised classification, applying a model acquired from the training set on the test set. Note, however, that the labeled training data are acquired automatically through the parallel corpus, without the need for human annotation.

Statistical Model. We train a Naive Bayes classifier, a simple but effective model for text categorization (Domingos and Pazzani, 1997). It predicts the class c for a sentence s by maximising the product of the probabilities for the features f given the class, multiplied by the class probability:

$$\hat{c} = \underset{c}{\operatorname{argmax}} P(c|s) = \underset{c}{\operatorname{argmax}} P(c)P(s|c) \quad (3)$$

$$= \underset{c}{\operatorname{argmax}} P(c) \prod_{f \in s} P(f|c) \quad (4)$$

We experiment with three sets of features. The first set consists of words, following the intuition that some words should be correlated with formal address (like titles), while others should indicate informal address (like first names). The second set consists of part of speech bigrams, to explore whether this more coarse-grained, but at the same time less sparse, information can support the T/V decision. The third set consists of one feature that represents a semantic class, namely a set of 25 archaic verbs and pronouns (like *hadst* or *thyself*), which we expect to correlate with old-fashioned T use. All features are computed by MLE with add-one smoothing as $P(f|c) = \frac{freq(f,c)+1}{freq(c)+1}$.

Results. Accuracies are shown in Table 3. A random baseline is at 50%, and the majority class (V) corresponds to 60%. The Naive Bayes models significantly outperform the frequency baselines at up to 67.0%; however, only the difference between the best (Words+Archaic) and the worst (Words+POS) model is significant according to a χ^2 test. Thus, POS features tend to hurt, and the archaic feature helps, even though it technically overcounts evidence.[5]

Model	Accuracy
Random BL	50.0
Frequency BL	60.1
Words	66.1
Words + POS	65.0
Words + Archaic	**67.0**
Human (no context)	64
Human (in context)	70

Table 3: NB classifier results for the T/V distinction

The Naive Bayes model notably performs at a roughly human level, better than human annotators on the same setup (no context sentences), but worse than humans that have more context at their disposal. Overall, however, the T/V distinction appears to be a fairly difficult one. An important part of the problem is the absence of strong indicators in many sentences, in particular short ones (cf. Example 1). In contrast to most text categorization tasks, there is no topical difference between the two categories: T and V can both co-occur with words from practically any domain.

Table 4, which lists the top ten words for T and V (ranked by the ratio of probabilities for the two classes), shows that among these indicators, many are furthermore names of persons from particular novels which are systematically addressed formally (like Phileas Fogg from Jules Vernes' *In eighty days around the world*) or informally (like Mowgli, Baloo, and Bagheera from Rudyard Kipling's *Jungle Book*).

Nevertheless, some features point towards more general patterns. In particular, we observe titles among the V-indicators (*gentlemen*, *madam*, *ma+'am*) as well as formulaic language (*Permit (me)*). Indicators for T seem to be much more general, with the expected exception of archaic *thou* forms.

5 Conclusions and Future Work

In this paper, we have reported on an ongoing study of the formal/informal (T/V) address distinction in

[5] We experimented with logistic regression models, but were unable to obtain better performance, probably because we introduced a frequency threshold to limit the feature set size.

Top 10 words for V		Top 10 words for T	
Word w	$\frac{P(w\|V)}{P(w\|T)}$	Word w	$\frac{P(w\|T)}{P(w\|V)}$
Fogg	49.7	Thee	67.2
Oswald	32.5	Trot	46.8
Ma	31.8	Bagheera	37.7
Gentlemen	25.2	Khan	34.7
Madam	24.2	Mowgli	33.2
Parfenovitch	23.2	Baloo	30.2
Monsieur	22.6	Sahib	30.2
Fix	22.5	Clare	29.7
Permit	22.5	didst	27.7
’am	22.4	Reinhard	27.2

Table 4: Words that are indicative for T or V

modern English, where it is not determined through pronoun choice or other overt means. We see this task as an instance of the general problem of recovering “hidden” information that is not expressed overtly.

We have created a parallel German-English corpus and have used the information provided by the German pronouns to induce T/V labels for English sentences. In a manual annotation study for English, annotators find the form of address very difficult to determine for individual sentences, but can draw this information from broader English discourse context. Since our annotators are not native speakers of English, but of languages that make the T/V distinction, we can conclude that English provides lexical cues that can be interpreted as to the form of address, but cannot speak to the question whether English speakers in fact have a concept of this distinction.

In a first statistical analysis, we found that lexical cues from the sentence can be used to predict the form of address automatically, although not yet on a very satisfactory level.

Our analyses suggest a number of directions for future research. On the technical level, we would like to apply a sequence model to account for the dependecies among sentences, and obtain more meaningful features for formal and informal address. In order to remove idiosyncratic features like names, we will only consider features that occur in several novels; furthermore, we will group words using distributional clustering methods (Clark, 2003) and predict T/V based on cluster probabilities.

The conceptually most promising direction, however, is the induction of social networks in such novels (Elson et al., 2010): Information on the social relationship between a speaker and an addressee should provide *global* constraints on all instances of communications between them, and predict the form of address much more reliably than word features can.

Acknowledgments

Manaal Faruqui has been partially supported by a Microsoft Research India Travel Grant.

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 the 26th Annual Meeting of the Association for Computational Linguistics*, pages 147–154, Buffalo, New York.

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.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 59–66, Budapest, Hungary.

Pedro Domingos and Michael J. Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2–3):103–130.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden.

Alexander Fraser. 2009. Experiments in morphosyntactic processing for translating to and from German. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 115–119, Athens, Greece.

Jerry Hobbs and Megumi Kameyama. 1990. Translation by abduction. In *Proceedings of the 13th International Conference on Computational Linguistics*, Helsinki, Finland.

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 the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1031–1040, Honolulu, Hawaii.

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

Michael Schiehlen. 1998. Learning tense translation from bilingual corpora. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, 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.

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.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics*, pages 200–207, Pittsburgh, PA.