International Journal of Computational Linguistics and Applications, Vol. 6, No. 1, 2015, pp. 11–24
Received 23/02/2015, Accepted 17/05/2015, Final 15/06/2015.
ISSN 0976-0962, http://ijcla.bahripublications.com

Identifying Linguistic Correlates of Social Power

RACHEL COTTERILL [1]
KATE MUIR [2]
ADAM JOINSON [2]
NIGEL DEWDNEY [1]

[1] *University of Sheffield, Sheffield, UK*
[2] *University of the West of England, Bristol, UK*

ABSTRACT

Previous work on social power modelling from linguistic cues has been limited by the range of available data. We introduce a new corpus of dialogues, elicited in a controlled experimental setting where participant roles were manipulated to generate a perceived difference in social power. Initial results demonstrate successful differentiation of upwards, downwards, and level communications, using a classifier built on a small set of stylistic features.

1. INTRODUCTION

One of the fastest growing areas of computational sociolinguistics in recent years has been the task of inferring various personal attributes from linguistic data. This is a popular mechanism for making sense of the social web and its ever-increasing quantities of data. Studies have spanned a range of topics, including classification by age, gender, native language, social group membership, and mental state.

The task of categorising relationships is a particularly interesting instance of the general problem. Unlike a demographic attribute such as age or native language, which is relatively stable for an individual across all communicative

contexts, we expect to see the same individual participating in a range of different social roles and relationships: the speaker's production is directly influenced by the specific audience.

The category of interpersonal relationships that has recieved the majority of scholarly attention to date is in the arena of hierarchy and social power, in part because this is a comparatively well-defined relation which is typically codified within an organisational structure. For example, managers are generally assumed to sit above their staff in the social hierarchy, professors are senior to students, and forum moderators have a position of power over ordinary contributors.

The majority of previous studies on categorising relationships and identifying power have relied on existing datasets such as Enron emails [1, 2], discussions between Wikipedia editors [3], and courtroom transcripts [4]. These studies have highlighted the shortage of publicly available datasets with high-quality ground truth. For example, Enron studies have made use of sparse hierarchies, reconstructed from publicly available information on organisational roles; these cover only a small subset of the individuals represented in the data, and do not form a well-connected graph [5, 6]. In the rare cases where experimental data has been gathered (e.g. [7]), these datasets have not been published, rendering them of limited use to the wider community.

This paper introduces a new, public dataset of transcribed speech, gathered in an experimentally controlled setting. We use this data to study the stylometric expression of social hierarchy.

2. PREVIOUS WORK

The effect of hierarchy and power on linguistic choices has always been of interest to linguists and sociologists. Brown & Levinson's [8] politeness theory identified relative power (the asymmetric relation) as one major factor of politeness in language, alongside social distance (the symmetric relation) and degree of imposition.

In more recent studies, computational approaches have examined qualitative approaches to large data sets. Peterson et al.

[5] investigate the applicability of Brown & Levinson's theory to email data, looking for correlations between informal features in text, and the level of politeness predicted by the theory. The features which they use to identify informal text include informal word lists, punctuation features (such as use of exclamation marks, or missing sentence-final punctuation), and case features (such as lowercase sentences). They report that informality features in the Enron email corpus are distributed largely as predicted by politeness theory.

Danescu-Niculescu-Mizil, et al. [3] study politeness within two online datasets: discussions between Wikipedia editors, and on Stack Exchange. They use Mechanical Turk to annotate turns with level of politeness, and demonstrate a distribution of politeness features in line with Brown & Levinson's predictions. They show that politeness is a precursor to promotion, at least in a community-approval model such as becoming an admin for Wikipedia: users who employ more politeness strategies are more likely to succeed in their social goals, and subsequently become less polite following promotion.

In another study of the Enron corpus, Bramsen et al [1] build an n-gram model and report a classification accuracy of 78.1% on the upspeak-downspeak task, and 44.4% accuracy on the three-way task of distinguishing upwards, downwards, and level communications. Cotterill [2] builds on Bramsen et al.'s work to model social power using only stylistic features, achieving comparable results with a smaller feature set.

Gilbert [9] examines the manifestation of power in the Enron corpus from a phrase-based perspective, using penalized logistic regression to identify those phrases which are particularly correlated with high or low power (as defined by job roles within the company). Using an SVM classifier to measure the predictivity of the resulting features, he reports an accuracy of 70.7% under three-fold cross validation.

Kacewicz et al. [7] undertake a series of five experiments with social power manipulation under different conditions, and report generalised findings relating to the differing use of pronouns. Lower-status individuals were observed to use more first person singular forms, while first person plural was used

more commonly by higher-status individuals. Second person forms were also used more by higher-status speakers, although the difference was less marked in this case.

3. DATA ELICITATION

We recorded and transcribed a collection of dyadic interactions as part of an applied psychology experiment into power-differential behaviour in a simulated business environment.

Volunteers were recruited from the student body at [anonymised] and given a task to complete, which they were advised concerned "creativity in business." A total of 41 participants took part in the study. The experimental group was composed of twelve participants assigned to the "judge" role and twelve "workers" (after [10]). The remaining 17 participants were assigned to the control condition.

In the experimental group the participants were randomly divided into judges and workers. The workers were given brief outlines of product ideas: these were drawn from Kickstarter campaigns, and featured an image and a short product description text. The workers pitched each idea to a judge, in a one-to-one conversation, and following a brief period of discussion the judges then chose whether or not to 'invest' in the concept. Both sets of participants were given to understand that the judges' ratings would affect the level of payment received by the workers for their participation, whereas the workers were given no such mechanism to provide feedback on the judges, thereby generating a scenario with a clear power differential between the two groups. (To satisfy the ethics board, eventual payment was in fact at a fixed rate for all participants.)

Members of the control group were similarly divided into two groups and provided with idea sheets, but instead of a worker/judge dynamic they were asked to discuss the inventions between themselves with an eye to potential collaborations. Neither party was given a higher status in the interaction, and they were informed that their participation payment would be a fixed amount, regardless of interaction success.

In both conditions, participants rotated through multiple conversation partners using a "speed dating" model to generate a number of independent one-to-one interactions lasting five minutes each. These exchanges were recorded, and after the end of the experiment the recordings were professionally transcribed. With a couple of exceptions due to corrupted files, one interaction was recorded between each judge/worker pair in the experimental condition (142 conversations) and between each pair in the control condition (72 conversations). The recorded conversations sum to 13,266 turns, giving a mean of 61.99 turns per dyad. The distribution of turns varied between the hierarchical ($\mu = 59.92$, $\sigma = 29.23$) and non-hierarchical ($\mu = 66.07$, $\sigma = 20.30$) condition, but this does not represent a statistically significant variation.

From a sociolinguistic perspective, the major disadvantage of this dataset is that it does not contain example utterances from the same individual participating under more than one role. A given student took on the role of judge, or worker, or part of the control group, and maintained this role for the duration of the experiment. It is therefore not possible to measure how an individual's linguistic choices shift in response to the changing of their relative power within a scenario.

A range of supplementary data was collected from each participant, including demographic information and personality profiling questionnaires. Most of the participants (82.9%) were undergraduate students from the University of [anonymised]. The remainder was made up of postgraduate students and non-students. Participants' ages ranged from 18 to 25. Female subjects made up 70.7% of the population, and 75.6% listed their ethnic origin as British.

As the data was elicited under controlled circumstances, we have reliable information concerning which participants were assigned to which social roles. The participants did not know one another in advance, so unlike in genuine organisational contexts, it is not necessary to account for the possibility of existing social relationships crossing these hierarchical boundaries in unexpected ways. As the roles were assigned at random, we also avoid the possibility of interference from underlying personality

traits or other demographic factors, which might lead to someone achieving a leadership role while also being expressed via their language choices.

With an experimental setup, there is always a risk that the participants' behaviour may be affected by the artificial nature of the setting. However, as we will demonstrate, the data still exhibits significant stylistic differences between speakers in different roles. After the experiment, a manipulation check was conducted by asking participants to score the level of power they felt they had during the interactions: results indicated that judges felt the most powerful ($\mu = 3.7$, $\sigma = 1.1$), while workers reported lower scores ($\mu = 2.8$, $\sigma = 1.1$), which is significantly different at 95%. Interestingly, both control groups rated their perceived power as less than either of the experimental groups ($\mu = 1.8$, $\sigma = 1.1$ and $\mu = 2.0$, $\sigma = 1$), which may be a consequence of participating in a scenario where their actions were not expected to change any of the outputs.

4. CLASSIFYING SOCIAL POWER

4.1. *Feature selection*

Following earlier work on social power modelling, we select a set of stylistic features to model our data. For email data, stylistic features have been shown to be broadly as effective as n-gram features, while resulting in a model of significantly lower dimensionality [2]. We apply an equivalent feature set, while noting that speech data lacks a number of the features that would be indicative of informality in text, such as varying capitalization or innovative punctuation.

One particular advantage of stylometrics is that selection of stylistic features tends to be subliminal: for example, in spontaneous production, an individual cannot control his use of function words such as pronouns or determiners.

A full list of features is included in Table 1. The majority of these are self-explanatory, but some would benefit from further elucidation.

Table 1. *List of stylometric features*

Characters per word	Interjections
Words per sentence	Expletives
Sentences per utterance	Contractions
Commas	Polite expressions
Periods	Hedging expressions
Semicolons	Deictic expressions
Colons	Modal verbs
Question marks	Verbs
Exclamation marks	Nouns
Hyphens	Pronouns
Parentheses	Determiners
Uppercase letters	Adjectives
Tag questions	Adverbs
Heylighen-Dewaele F-score	Prepositions
Out-of-vocabulary words	Conjunctions
Numbers	

Because the data has been professionally transcribed, there is less chance of typographical errors, contrasted with text that has been spontaneously produced – and if such errors do exist, they are due to the transcriber rather than the participant. Nevertheless, a measure of out of vocabulary words (measured with respect to an English dictionary) may prove a valuable feature as this encompasses a number of phenomena including codeswitching, informal slang, and highly technical jargon.

We retain the distribution of punctuation as a feature set, on the assumption that the transcriber's selection of punctuation will reflect speech-related features such as timing and pitch. Similarly, the concept of a 'sentence' in speech is controversial, but we nevertheless retain it as a feature for comparison with earlier work. The distribution of uppercase letters is also employed as a useful proxy for proper nouns (encompassing some such as product names which may not be captured by an entity tagger).

Parts of speech are tagged using the OpenNLP toolkit. Heylighen and Dewaele's F-score [11] is a linear combination of parts of speech, following a formal definition of contextuality; this is included as a separate feature.

4.2. *Individual message results*

A random forest classifier (using WEKA) was trained over the stylistic features from Table 1, and performance was assessed using five-fold cross validation. The logical baselines for this task are the random baseline, at 33.3%, and the most common class (level) 35.86%.

Message-level accuracy was 41.98%, using all features. Broken down further, this represents 36.59% accuracy for messages going up the hierarchy, 40.79% for downwards messages, and 47.87% accuracy for messages that formed part of peer-level exchanges.

From the resulting confusion matrix it is evident that level communications are the most successfully classified, but at the cost of classifying a number of upwards and downwards messages into the 'level' category.

Table 2. *Confusion matrix: message level results. Columns are predicted values, rows are truth*

	Upwards	Downwards	Level
Upwards	**1556 (11.7%)**	1144 (8.6%)	1553 (11.7%)
Downwards	1081 (8.1%)	**1736 (13.1%)**	1439 (10.8%)
Level	1217 (9.2%)	1263 (9.5%)	**2277 (17.2%)**

It is also interesting to consider individual variation. Classification accuracy at the individual level (calculated across all messages sent by that individual) ranges between 16.6% and 69.3%, following an approximately normal distribution (μ = 42.7, σ = 11.0). From this we can see that some individuals use language in a way that is 'more typical' of their role, while others are more divergent in their linguistic behaviour.

Our results at this stage are above baseline performance, although a couple of percentage points below the results reported for email in [1] and [2]. This is clearly unlikely to represent sufficient performance for any real-world applications, so we will proceed to examine ways in which accuracy can be enhanced.

4.3. *Simple plurality voting*

So far, we have considered categorisation at the message level, with results that are promising but not groundbreaking. However,

it is unlikely that any individual message perfectly captures the entire essence of a pair's relationship, and as such, we might expect to get better results by combining predictions from multiple messages.

There are two distinct methods for approaching such a task: a classifier can be trained on the aggregate features of the whole message set, or the results of single-message classification can be combined in an additional step. Since we have already obtained above-chance performance at the single-message level, we adopt the second approach.

The most basic method of combining scores is to use a 'voting' method. For example, given a set of twenty messages between A and B, we might have the following output from our individual message classification:

A to B	upwards: 7	downwards: 2	level: 1
B to A	upwards: 4	downwards: 6	level: 0

Based on these numbers, we would have one vote for an equal relationship, and 19 for a hierarchy. Looking further into the hierarchical evidence, we find 7 + 6 = 13 votes for A being subordinate to B, and 2 + 4 = 6 votes for B being subordinate to A. In this case, if A is indeed B's subordinate, we have the potential to turn 65% message-level accuracy into a single correct prediction at the relationship level. Of course, the inverse of this is that when we get it wrong, we will be degrading our overall performance.

For our initial experiments with aggregation, we simply took as our answer whichever case had the highest number of votes in total. We will refer to this technique as 'simple plurality voting', by analogy with electoral systems such as first-past-the-post.

Considering aggregation at the level of the individual speaker, applying simple plurality voting to the classifier output gives us two predictions for each dyad, one based on each speaker's output. We assess the accuracy of these predictions independently, and observe that our overall mean accuracy increases to 57.9% at the speaker level — but variance also increases, from 11.0 to 27.0, and our distribution is no longer

normal. Figure 1 demonstrates this shift in distribution. Note that we have 41 speakers producing 13,266 turns: we display results as percentages for ease of comparison, but in absolute terms, all numbers are much smaller in the aggregated case.

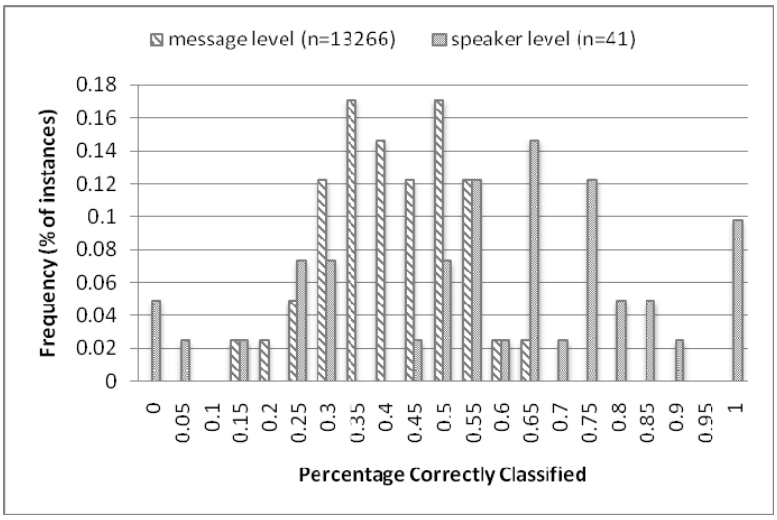

Figure 1. *Chart demonstrating the distribution of correctly-classified instances for individual messages, and for speaker-level aggregation.*

Speaker-level aggregation is simple and informative, but still leaves us with two predictions for each dyadic relationship, which may be in conflict. We extend the simple plurality voting method to include votes in both directions, as a single system set up to generate one prediction per pair. Due to the collection methodology of the experimental dataset we are guaranteed, for each dyad, an approximately equal number of utterances in each direction, so in this instance there is no need to concern ourselves with imbalances in the data.

Using simple plurality voting on a pairwise basis we achieve 69.1% accuracy in the task of three-way prediction across pairs, with seven pairs unassigned (cases where there was no single 'most common' class).

The resulting error analysis shows that the system is more likely to mis-categorise relationships as level when they are actually hierarchical; by comparison, incorrectly inverting the hierarchy is relatively rare.

Table 3. *Error analysis: Pairwise aggregation (correct lines in bold). Percentages exclude the seven uncategorised instances*

82	**39.61%**	**Hierarchical, correctly labelled**
45	21.74%	Hierarchical, incorrectly labelled as level
10	4.83%	Hierarchical, labelled with incorrect polarity
61	**29.47%**	**Level, correctly labelled**
9	4.35%	Level, incorrectly labelled as hierarchical

4.4. *The effect of thresholds*

Intuition suggests that it should be possible to obtain a higher degree of accuracy by setting a minimum confidence threshold, and accepting classifications only above this threshold.

One simple method of applying a threshold to a voting system is to set a minimum percentage of messages which must fall into the 'most popular' classification before it can be accepted. For a three-class problem such as this, the default (and lowest possible) threshold for a simple plurality vote is 0.33, as it isn't possible for all three classes to obtain less than a third of the available votes.

We investigated setting higher thresholds, from 0.4 up to 0.6. Increasing the threshold gives an almost linear improvement in raw accuracy (over the classified instances), but at the cost of rejecting ever higher number of instances without classification. The improvement in precision comes with a fairly steep drop in recall once the threshold is above 0.4. As always, the appropriate compromise between precision and recall will vary depending on the application.

Table 4. *Effect of thresholding on precision and recall*

	0.333	**0.4**	**0.5**	**0.6**
Unclassified pairs	7	25	109	177
Accuracy (%)	69.08	71.96	75.24	78.37

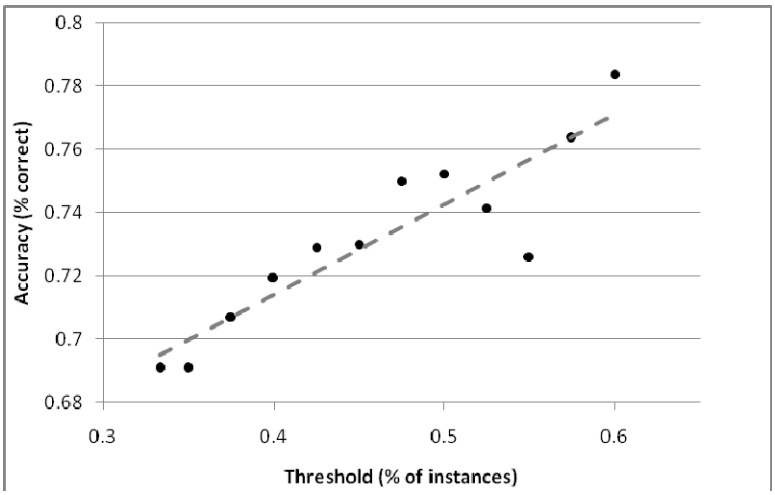

Figure 2. *Trend of accuracy as threshold is raised*

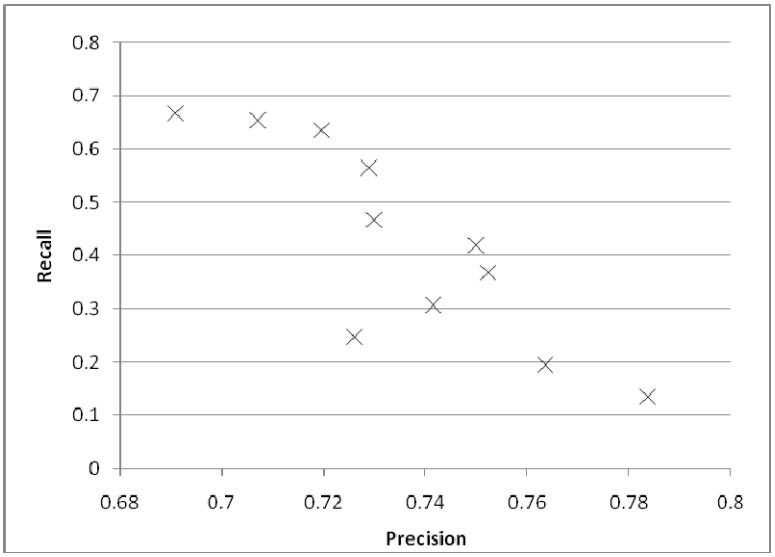

Figure 3. *Precision-recall plot for varying confidence thresholds*

5. Conclusions and Future Work

Using a small set of stylistic features, we have achieved above-chance performance at the individual message level for classifying spoken dialogues. Additionally, we have demonstrated a significant improvement in performance as a direct result of aggregating data at the relationship level. We have shown that introducing a threshold can improve precision, but only at the cost of a significant drop in recall, which is unlikely to be a worthwhile trade-off in real world applications.

In future work we intend to address a number of limitations of our experimental set-up. We plan to replicate our data collection step using a computer-mediated setting, to allow for direct comparison of spoken and textual conversations. Additionally, we hope to design a suitable scenario which would allow for the possibility of participants participating under more than one role.

References

1. Bramsen, P., Escobar-Molano, M., Patel, A. & Alonso, R. 2011. Extracting social power relationships from natural language. In proceedings of the *Annual Meeting on Association for Computational Linguistics* (pp. 773-782).
2. Cotterill, R. 2013. Using stylistic features for social power modeling (El uso de características estilísticas para modelado del poder social). *Computación y Sistemas*, 17/2, 219-227.
3. Danescu-Niculescu-Mizil, C., Sudhof, M., Leskovec, J. & Potts, C. 2013. A computational approach to politeness with application to social factors. In *Proceedings of ACL*.
4. Danescu-Niculescu-Mizil, C., Lee, L., Pang, B. & Kleinberg, J. 2012. Echoes of power: Language effects and power differences in social interaction. In proceedings of the *21st International Conference on World Wide Web* (pp. 699-708), New York, NY, USA.
5. Peterson, K., Hohensee, M. & Xia, F. 2011. Email formality in the workplace: A case study on the Enron corpus. In proceedings of the *Workshop on Languages in Social Media* (pp. 86-95).

6. Agarwal, A., Omuya, A., Harnly, A. & Rambow, O. 2012. A comprehensive gold standard for the Enron organizational hierarchy. In proceedings of the *50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2* (pp. 161-165), Stroudsburg, PA, USA.
7. Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M. & Graesser, A. C. 2013. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, Sep.
8. Brown, P. & Levinson, S. C. 1987. *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.
9. Gilbert, E. 2012. Phrases that signal workplace hierarchy. In proceedings of the *ACM 2012 Conference on Computer Supported Cooperative Work* (pp. 1037-1046), New York, NY, USA.
10. Guinote, A., Judd, C. M. & Brauer, M. 2002. Effects of power on perceived and objective group variability: Evidence that more powerful groups are more variable. *Journal of Personality and Social Psychology*, 82/5, 708-721.
11. Heylighen, F. & Dewaele, J. M. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7, 293-340.

RACHEL COTTERILL
UNIVERSITY OF SHEFFIELD, SHEFFIELD, UK.
E-MAIL: < R.COTTERILL@SHEFFIELD.AC.UK>

KATE MUIR
UNIVERSITY OF THE WEST OF ENGLAND, BRISTOL, UK.
E-MAIL: < KATE.MUIR@UWE.AC.UK>

ADAM JOINSON
UNIVERSITY OF THE WEST OF ENGLAND, BRISTOL, UK.
E-MAIL: < ACP08NJD@SHEFFIELD.AC.UK>

NIGEL DEWDNEY
UNIVERSITY OF SHEFFIELD, SHEFFIELD, UK.
E-MAIL: < ADAM.JOINSON@UWE.AC.UK>