Finding Zelig in Text: A Measure for Normalizing Linguistic Accommodation

Simon Jones[1], **Rachel Cotterill**[2], **Nigel Dewdney**[2], **Kate Muir**[3], **and Adam Joinson**[3]

[1]Department of Computer Science, University of Bath, BA2 7AY. s.jones@bath.ac.uk
[2]Department of Computer Science, University of Sheffield, S1 4DP. r.cotterill@sheffield.ac.uk, acp08njd@sheffield.ac.uk
[3]Behavioural Research Lab, Faculty of Business and Law, University of the West of England, Bristol, BS16 1QY. kate.muir@uwe.ac.uk, adam.joinson@uwe.ac.uk

Abstract

Linguistic accommodation is a recognised indicator of social power and social distance. However, different individuals will vary their language to different degrees, and only a portion of this variance will be due to accommodation. This paper presents the *Zelig Quotient*, a method of normalizing linguistic variation towards a particular individual, using an author's other communications as a baseline, thence to derive a method for identifying accommodation-induced variation with statistical significance. This work provides a platform for future efforts towards examining the importance of such phenomena in large communications datasets.

1 Introduction

"Zelig...protects himself by becoming like whoever he is around."
- The Narrator, Zelig (Allen, 1983)

When people converse, they often become more alike in their language in many different dimensions (Garrod and Pickering, 2004). This can include similarity in pronunciation (Giles, 1973), speech rates (Street, 1984), pause and utterance duration (Cappella, 1979), and volume (Natale, 1975). Similarly, in written communications people often converge in terms of features such as linguistic style (Danescu-Niculescu-Mizil and Lee, 2011), vocabulary and syntax (Scissors et al., 2008). Communication Accommodation Theory (Giles and Ogay, 2007) proposes that interactants can adjust their communication style, such as accent, vocabulary, and use of jargon to sound more (convergence) or less (divergence) like the other person. Individuals typically converge to signal affinity with their interlocutor, and diverge to show interpersonal or social distance.

One area which has been largely overlooked, to date, is the role played by an individual's inherent tendency to accommodate (or not). We propose that some people are more apt than others to change their typical linguistic style to converge to that of their conversational partner. This paper introduces the Zelig Quotient, which is a new method for capturing the degree to which the variation in an individual's language use can be explained by their accommodation towards the style of their interlocutor. Using this score, it is then possible to measure the significance of an individual's accommodation within a specific communication pair: does each individual accommodate more or less than their personal norm?

In this paper, we firstly consider existing computational measures of linguistic accommodation. Although useful in measuring accommodation of specific linguistic features within dialog, current measures

do not permit examination of the role of an individual's inherent (or latent) predisposition to accommodate their linguistic style. Further, another area that has yet to be explored is the influence of social status and relationships between interlocutors on the likelihood of accommodation. We therefore demonstrate the applicability of the Zelig Quotient by applying the technique to large datasets of communications from three online community forums, in which social status and relationships between interlocutors are clearly defined. We close with a discussion of potential future directions in which the Zelig Quotient could be applied.

1.1 Computational Measures of Linguistic Accommodation

Several computational measures of linguistic accommodation already exist. These measures typically capture the extent to which language use increases in similarity or becomes 'adapted' either within a piece of text or between individuals in dialog. Church (2000) developed a method for determining lexical adaptation in text, by examining the probability of one word appearing in the later half of a document when it appears in the earlier half. This method has been used and extended by other researchers in the examination of lexical adaptation over time (Reitter et al., 2006), adaptation of syntactic constructions (Dubey et al., 2005) and to measure the prevalence and strength of linguistic feature adaptation in dialogs (Stenchikova and Stent, 2007). Along similar lines, linguistic style matching (LSM) techniques with LIWC measures (quantitative analysis of standard language categories) (Pennebaker et al., 2001) reveal the extent to which language use is coordinated between group members, either on a whole conversation or turn-by-turn level (Niederhoffer and Pennebaker, 2002). Ireland et al. (Ireland et al., 2011) used LSM techniques to study the predictive value of stylistic similarity, in a social setting. They found that similarity of a few stylistic categories (such as the distribution of pronouns and determiners) was a good indicator of whether two individuals would vote to see one another again in a speed dating scenario.

The work of Huffaker et al. (2006) is particularly relevant to our examination of accommodation within online community forums. Huffaker et al compared three different measures of lexical convergence to assess message similarly in an online community over time. These included: Spearman's Rank Correlation, which has been used to determine message similarity between corpora (Kilgarriff, 2001); 'Zipping', referring to data compression algorithms, which has been used to measure the complexity of documents (Benedetto et al., 2002); and Latent Semantic Analysis, which has been used to measure semantic similarity across corpora (Coccaro and Jurafsky, 1998). All three measures showed divergence in message similarity both between individuals, and in the community as a whole across time.

However, the common theme with all these techniques is that although they can effectively measure adaptation of linguistic feature use within and between dialogs, they fail to capture the precise direction of convergence or divergence between individuals (i.e., do both interactants within a conversational pair accommodate their language use to the same extent?) Thus, existing computational measures of linguistic accommodation fail to provide a fine-grained view of the dynamics of convergence within dyads or large groups. The measures discussed here only capture the extent to which members of the group match one another, and overlook precise details of individuals movements from their existing language use towards that of the group.

1.2 Individual's Propensity to Accommodate

Further, whilst accommodation within dyads and groups has been measured extensively, one area which has been largely overlooked, to date, is the role played by an individual's inherent tendency to accommodate (or not). Some individuals may have a relatively stable linguistic style, whereas other individuals may be more likely to accommodate their linguistic style towards that of their conversational partner. We have been unable to find any methods for reliably measuring an individuals propensity to accommodate towards their interlocutors. We hypothesize that individuals are not equal with respect to their accommodation and propose the Zelig measure, detailed within this paper, as a means for quantifying this characteristic.

One factor which could conceivably influence an individual's tendency to accommodate is *social power*. Giles and Coupland (1991) state that "the power variable is one that emerges a number of times

Category	Examples	Category	Examples
Personal pronouns	I, his, their	Auxillary verbs	shall, be, was
Impersonal pronouns	it, that, anything	High-frequency adverbs	very, rather, just
Articles	a, an, the	Negations	no, not, never
Conjunctions	and, but, because	Quantifiers	much, few, lots
Prepositions	in, under, about		

Table 1: Word Categories Used for Calculating Linguistic Style

in the accommodation literatures and in ways that support the model's central predictions". Demonstrations of the role of social power in accommodation include interviewees converging their speech style towards that of their interviewers during employment interviews (Willemyns et al., 1997), students accommodating their verbal and non-verbal behaviours to academic faculty members (Jones et al., 1999) and witnesses in courtrooms accommodating to the linguistic style of the questioning legal professional (Gnisci, 2005). Thus, individuals with low social power are more likely to accommodate their linguistic style. The Zelig Quotient allows explicit examination of research questions of this nature concerning accommodation and divergence associated with demographic variables such as social power.

2 The Zelig Quotient

"Wanting only to be liked, he distorted himself beyond measure."
- The Narrator, Zelig (Allen, 1983)

We propose the Zelig Quotient, a measure for normalizing linguistic variation. The Zelig Quotient is named for Leonard Zelig, the central character of the Woody Allen film Zelig, who is described as "the human chameleon" due to his propensity for taking on the characteristics of other people. This is the logical extreme of accommodating to one's audience. An author who always adopts the language style of the intended reader is totally Zelig-like, whereas an author who does not adapt at all has zero likeness. Opposite behaviour to Zelig (moving away from the audience) is also possible. Over-accommodation occurs when the author adopts elements of linguistic style of their intended reader, but emphasises to the point of overuse. In extreme cases this would be detected as parody. We need, therefore, to distinguish not only the distance between author and reader, but also the orientation. The Zelig Quotient thus shows the extent to which an individual changes their linguistic style from their 'typical' or baseline style, to move either towards or away from each of their conversational partners. The average Zelig score across all conversational partners can then be used to demonstrate the individual's general tendency to accommodate their language use to that of others.

2.1 Feature Selection

We have selected to study a set of features which are stylistic rather than semantic in nature; although consideration of whether two people are talking about the same topic is a valid research question, we currently wish to focus on their linguistic style. The best features for our purposes are those able to be varied with comparative freedom, without affecting the meaning of a message. We use a set of nine such features, taken from the linguistic style matching study conducted by Ireland et al. (2011) (see Table 1).

We used LIWC dictionaries for each category. LIWC processes a text file word by word, comparing each word to the dictionary and providing a count of the words in the file which match each category in the dictionary. Sums of words in each category are presented as percentage of total words in the file to correct for differences in text length between text files (Pennebaker et al., 2001). The use of LIWC is the basis of much recent work on linguistic style accommodation (Niederhoffer and Pennebaker, 2002; Danescu-Niculescu-Mizil et al., 2011; Danescu-Niculescu-Mizil and Lee, 2011) to which we want to relate.

2.2 Calculating the Zelig Quotient

We assume an author has a baseline linguistic style resulting in a baseline value for each of our stylistic features. However, we expect variation in the observed values due to sampling, an author's natural variation, constraints of message content and format, etc. as well as any movement due to accommodation.

We can estimate a baseline value for a specific feature, μ_f, by averaging over all the messages we have for an author a. Previous research has used a similar technique for establishing the baseline level of a lexical item in a dialog in order to study accommodation (Church, 2000; Stenchikova and Stent, 2007).

$$\mu_f(a) = \sum_{m=1}^{n_a} f_m(a)/n_a \tag{1}$$

where m is a message, n_a is the number of messages for a and $f_m(a)$ is the feature value in m.

We can further estimate the proportion of variance due to 'noise' and that due to accommodation by also calculating the average feature values on an author-reader (a, r) pairwise basis.

$$f(a, r) = \sum_{m=1}^{n_{ar}} f_m(a, r)/n_{ar} \tag{2}$$

where n_{ar} is the number of messages written by a to reader r.

Measuring the variance within a pair and then averaging over all pairs that author is party to gives an estimate of the natural variation in feature value for an author.

$$\sigma_f^2(a) = \frac{1}{R_a} \sum_r \sum_{m=1}^{n_{ar}} (f_m(a) - \mu_f(a))^2/n_{ar} \tag{3}$$

where R_a is the number of recipients of messages from author a.

The movement in a feature due to accommodation is simply taken to be the difference between the value seen within a communicative pair, and the author's baseline value.

Having calculated scores for several features, some of which may change more readily than others, we can consider authors as having corresponding points in an F-dimensional feature space described by the vector of feature values. The generalised phenomenon of accommodation can then be measured in terms of movement in this feature space, rather than movement in individual features. Note that to avoid bias towards particular features when considering overall movement, feature scales must be comparable. Movement in an author's language may be large, but it may not necessarily be towards the reader.

We represent movement and distances between the author's baseline position, accommodated position, and the reader's position as vectors in the feature space:

$$\mu = \{\mu_1(a), \mu_2(a), ..., \mu_F(a)\}$$
$$a = \{f_1(ar), f_2(a, r), ..., f_F(a, r)\}$$
$$r = \{f_1(r), f_2(r), ..., f_F(r)\}$$

We use the law of cosines to yield the cosine of the angle between the vector connecting the reader to the author's baseline position, and that connecting the reader to the author's accommodated position. The angle will be greater than 90° if the author has over-accommodated, and will therefore have a negative cosine value. However, the dot product of these two vectors gives the cosine of the inner angle. Therefore, normalising by this gives a value of +/- 1 according to whether accommodation movement is less or more than the amount required to meet the reader.

Multiplying the accommodated distance by this +/-1 factor gives us a definition of an accommodation metric that expresses the accommodation as the change in directed distance from the reader, proportional to the amount required from the author's unaccommodated position. This may be greater than 1 (over-accommodation) or less than zero (divergence). In vector notation we define accommodation as:

$$Acc(a, r) = 1 - \left(\frac{|\vec{ar}|}{|\vec{\mu r}|}\right)\left(\frac{|\vec{\mu r}|^2 + |\vec{ar}|^2 - |\vec{\mu a}|^2}{2(\vec{\mu r} \cdot \vec{ar})}\right) \tag{4}$$

The dot product of $\vec{\mu r}$ and $\vec{ar}$ is zero if the two vectors are orthogonal. However this is matched by a zero value in the numerator and we take the final parentheses value in equation (4) to be 1 in this case. In the other pathological case where $|\vec{\mu r}|$ is zero, the implication is that author and reader have the same preferred position, i.e. there is nothing meaningful to say about accommodation between the two.

Having estimated author to reader accommodation, we are now in a position to estimate how readily the author adapts to others, by averaging over the set of readers. This gives us our Zelig factor, Z.

$$Zelig(a) = \frac{1}{R_a}\sum_{r=1}^{R_a} 1 - \left(\frac{|\vec{ar}|}{|\vec{\mu r}|}\right)\left(\frac{|\vec{\mu r}|^2 + |\vec{ar}|^2 - |\vec{\mu a}|^2}{2(\vec{\mu r}\cdot\vec{ar})}\right) \tag{5}$$

A positive (+) Zelig Quotient signifies the author readily accommodates, with a Zelig Quotient of 1 indicating the author always adapts their linguistic style to that of their audience. A negative (-) Zelig Quotient suggests divergence in the authors linguistic style (moving away from the audience).

Significance of values can be estimated from the variance. Here we take movement beyond one standard deviation of the authors total message distribution. The significance of an author's Zelig Quotient then follows from averaging the variance seen over the pairs the author is party to.

$$Zelig_{min}(a) = \sqrt{\frac{1}{R_a}\sum_{r=1}^{R_a}\frac{\sum_{f=1}^{F}(\sigma_f(a) - \mu_f(a))^2}{\sum_{f=1}^{F}(f(r) - \mu_f(a))^2}} \tag{6}$$

This model assumes there are latent baseline distributions for feature values but does not suggest a generative function. Further work will determine appropriate distribution models for features, to be parameterised from the estimation methods presented here.

3 Zelig in Online Communications

To demonstrate the utility of the Zelig Quotient, our study uses scraped forum data from three large online communities (note, the names of the forums have been anonymised to protect the identity of the community members).

The first (ForumA) contains circa 2800 threads, 21000 posts, 250 currently active members and three years of historical data. The second (ForumB) contains approximately 160,000 threads, 2.25 million posts, 1500 currently active members and historical data is available for a period of approximately 10 years. The third (ForumC) contains approximately 50,000 threads, 550,000 posts, and currently 824 active members. Historical data is available for a period of approximately seven years.

All three of the online community forums are powered by vBulletin, a system which allows users to earn reputation points for their activity. Users can 'up-vote' or 'down-vote' each others' posts, which either adds or subtracts reputation points from that user. The number of reputation points received or deducted depends on who is casting their up/down-vote. Having more reputation enables a voter to have a greater influence on the reputation of others. Reputation can also be earned as the number of posts made by a user, or the age of their account, increases. This system essentially enables a power structure within the community, and is useful for differentiating between veteran communities members who contribute a lot to the community and, based on their up-votes, have a considerable amount of expertise or valuable information/opinions to share (i.e. Leaders), from relatively new and inexperienced community members whose contributions are less significant (i.e. Non-Leaders).

From each community we sample the top 10% of all members from the complete historical data (ForumA n = 70, ForumB n = 98, ForumC n = 169) based on reputation score and assign them to our 'Leader' category. For our 'Non-Leader' category we select an equally sized sample group (same n), which are evenly distributed across the remaining 90%, based on reputation score. One-time posters were removed prior to sampling.

3.1 Hypotheses

We hypothesise that, in accordance with Communication Accommodation Theory (Giles and Ogay, 2007), the Zelig Quotients for high power individuals (which we will refer to as Leaders) and low power

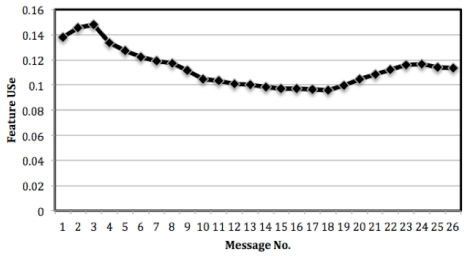


Figure 1: Moving average for an individual's linguistic feature use

individuals (referred to as Non-Leaders) will differ significantly. The power variable has been shown to have a strong influence on communication accommodation. Since people with lower status have greater cause to try to gain social approval by converging towards others, we hypothesise that:

H1: Non-Leaders (i.e. people with low power) will exhibit greater linguistic style accommodation/more Zelig-like behaviour than Leaders (people with high power)

Furthermore, it has been shown that those with low power often show greater convergence when communicating with somebody in a superior position. For example, foremen converge more to managers than to workers, and managers converge more to higher managers than to foremen (Taylor et al., 1978). Similarly, salespeople converge more to customers than vice versa, as the customers in these settings hold greater economic power (Van den Berg, 1986).

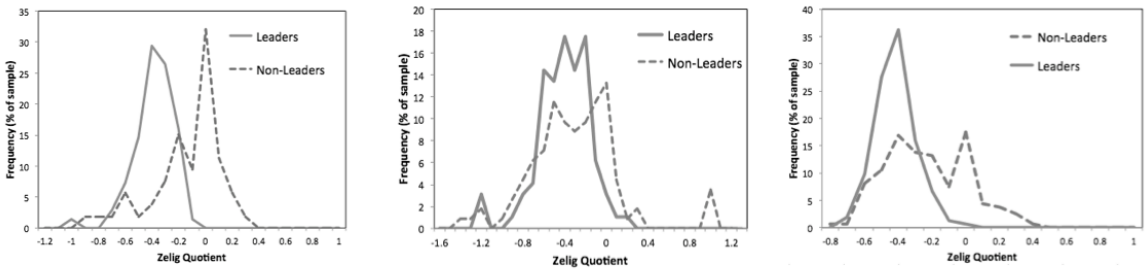
H2: People accommodate more with interlocutors who have higher power (Leaders) than those who have lower power (Non-Leaders)

3.2 Method

One challenge when working with scraped data from online community discussion threads, is accurately reconstructing who is talking to whom. Unlike in e-mail communication, vBulletin forums lack a mechanism for explicitly stating who a post is a reply to. Posters therefore append their post to the end of an ever-growing thread, regardless of whether they are addressing the first post, last post or any post in between. Of course, their communication may not even be aimed at any single person, and instead intended for a whole community audience.

Since the Zelig measure we have presented requires dyadic comparisons of linguistic style features, it is necessary to reconstruct a dyadic conversation structure for all of the forum threads. Previous work has examined features for accurate reply reconstruction of threaded conversations (Aumayr et al., 2011); re-building the correct structure from a collapsed conversation thread without explicit reply mechanisms. Many features are useful for reply graph reconstruction, for example: reply distance (how closely a post appears to that which it is responding), time difference (how soon after a post a response is written), quotation links (how explicit citations of previous posts are used) and cosine similarity (how closely the contents of two posts match). Aumayr et al. (2011) demonstrated that accuracy (as indicated by measurements of precision and recall) can be achieved by simply combining the use of reply distance and quotation links. That is to say, posts are typically responses to those which they appear closest to within a thread, or those which they explicitly cite. Therefore, for our analysis we treat each post as a response to the author of the closest post (the one directly preceding it within the thread) or the author that is cited within the post.

In order to calculate variations in each individuals' linguistic style, we calculate their baseline style as the moving average of their previous communications (either globally or within each particular dyad). Figure 1 illustrates this moving average for a particular LIWC feature changing with each message sent by an individual. Our moving average approach has the advantage that accommodation is calculated according to the movement towards a persons' linguistic style at a given point in time, rather than simply an average of their linguistic style in their entire communications (including those that occur later).



(a) Forum A Distributions (b) Forum B Distributions (c) Forum C Distributions

Figure 2: Zelig Quotient Distributions for members of Forum A, Forum B and Forum C

4 Results

Figure 2 shows the frequency distributions of Zelig Quotients for individuals within each of the three communities.

Our results show that the Zelig Quotients of community members follow a relatively normal distribution, centered around a Zelig Quotient of approximately -0.4 in all three communities, though there are obvious differences between the Zelig Quotient distributions for our two sample groups of Leaders vs. Non-Leaders.

Our analysis reveals that divergent communication is a common behaviour for a large proportion of each online community. That is, many community members receive a negative Zelig Quotient. Our results, therefore, go against the prevailing findings in linguistics and psychology, which suggest that individuals often constitute themselves as a community, speaking in a collective voice, and converging in terms of linguistic style.

4.1 Divergence is common

Within communication accommodation theory, convergence is generally regarded as positive and divergence as negative. Divergent communicators are often evaluated as insulting, impolite, and hostile (Bradac and Giles, 2005). Convergent speakers are evaluated as more competent, attractive, likeable and cooperative (Giles et al., 1991). Divergence is the typically the result of communicators wanting to differentiate themselves from each other and emphasize distinct identities. This can be the case particularly where there are power or status differences between interlocutors, as individuals attempt to communicate their social differences by engaging in dissimilar communication behaviours (Street, 1991). If their points of view start deviating, so to do their communication styles (McPherson et al., 2001). Observing the Zelig distribution of a community is therefore likely to provide a valuable insight into the overall 'unity' of its' members. Our results are consistent with Huffaker et al. (2006), who found increasing dissimilarity with the words used by community forum users over a six-week period. Thus, our results suggest that a large proportion of individuals within online communities have a tendency to differentiate themselves from others.

However, divergent communication is not always inherently negative. Attributions of the speaker's motives by the recipient can influence the extent to which convergent and divergent communications are perceived to be positive or negative. For instance, convergence can be evaluated as positive when attributed to speakers internal positive motives; however convergence can also be viewed negatively when attributed to external factors. The same is true for divergent communications. When divergence is perceived by the recipient as being unintentional or positively motivated, recipients evaluate the speaker and their communications more favourably than if it is evaluated as being intentional or negatively motivated (Gasiorek and Giles, 2012). Thus, divergent communication does not neccessarily have negative connotations for the relationship between speakers and recipients.

Whilst linguistic style convergence does occur within the communities we have examined, a relatively small proportion of members (typically less than 10%) demonstrate a tendency to accommodate to others

in a near Zelig-like way. A larger proportion of each community (between 15 - 35%) tend to maintain their typical linguistic style, with their Zelig Quotient of 0 indicating that fluctuations in style are due to noise rather than convergence or divergence.

The following results describe in more detail the differences in Zelig Quotients for our two sample groups, Leaders and Non-Leaders.

4.2 Non-Leaders are more Zelig-like than Leaders

We conducted independent samples t-tests in order to compare the Zelig Quotient means for Leaders and Non-Leaders in each of the online communities. There were common significant differences across all three communities: Non-Leaders in ForumA had significantly greater Zelig Quotients (M= -0.181, SD= 0.270) than Leaders (M= -0.432, SD= 0.148); t(119)= 6.492, $p < 0.001$. Similarly, ForumB Non-Leaders (M= -0.345, SD= 0.425) were more Zelig-like than Leaders (M= -0.461, SD= 0.254); t(206)= 2.346, $p < 0.05$, and ForumC Non-Leaders (M= -0.306, SD= 0.276), were also more Zelig-like than Leaders (M= -0.468, SD= 0.121); t(327)= 6.885, $p < 0.001$.

These results lead us to accept H1; Non-Leaders linguistic style variation can be attributed to their accommodation towards the style of their interlocutor, to a greater degree than for Leaders. Furthermore, our results illustrate that analysis using the Zelig Quotient uncovers important accommodation trends within textual conversation data, which are in accordance with Communication Accommodation Theory.

As well as a comparison between high and low reputation groups (Leaders and Non-Leaders, respectively), a Spearman's Rank Order correlation was run to determine the relationship between reputation rank and Zelig rank within the community. We found weak but statistically significant negative correlations between reputation and Zelig Quotient within ForumB (rs(316) = -0.1406, p = 0.011889) and ForumC (rs(179) = -0.1729, p = 0.019863), further suggesting that Zelig-like behaviour is most common within the lowest reputation ranks.

4.3 Interactions between Leaders and Non-Leaders

In order to test H2 and examine how interactions between Leaders and Non-Leaders influenced accommodation, each individual within a dyad was classified as either communicating 'Up' or 'Down' the reputation hierarchy (either from Non-leader to Leader, or Leader to Non-Leader, respectively). Independent samples t-tests were conducted in order to compare the mean Zelig Quotients for dyads from each category. The tests revealed a statistically significant difference within two of the three communities, with Zelig Quotients significantly lower in 'downward' communications for ForumA - Upwards (n = 72, M= -0.283 , SD= 0.456), Downwards (n = 72, M= -0.588, SD= 0.711); t(142)= 3.066, $p < 0.01$, and ForumC - Upwards (n=1280, M= -0.388, SD= 0.570), Downwards (n=1280, M= -0.530, SD= 0.723); t(2558)= 5.528, $p < 0.001$. Within ForumB, those communicating up the hierarchy were also more Zelig-like, however the difference was considered not to be statistically significant - Upwards (n=295, M= -0.383, SD= 0.634), Downwards (n=295, M= -0.444, SD= 0.601); t(588)= 1.208, $p = 0.228$.

These results lead us to accept H2; community members are more Zelig-like when communicating with people above them in terms of reputation/status.

4.4 Finding the Zeligs: Who are they?

To conclude our analysis, we address the question: 'Who are the Zelig characters within our corpora?'. By focusing our attention on the individuals that have positive $Zelig$ values ($Z > 0$), our results point to a clear and consistent answer across all three datasets - *almost all Zelig-like individuals are Non-Leaders*, however, Non-Leaders are *not* all Zeligs. Within our sample populations for each community Non-Leaders account for the vast majority of those with Zelig values greater than 0 (100% in ForumA, 91.6% in ForumB, and 100% in ForumC). These results are consistent with the idea that individuals with lower power are sensitive to the language used by a higher power interlocutor (Niederhoffer and Pennebaker, 2002); the Zelig Quotient has captured the greater tendency of Non-Leaders to alter their baseline linguistic style in order to converge with the linguistic style of Leaders, instead of the other way around.

The Zelig-like behaviour of non-leaders could potentially be attributable to their acclimatisation to community expectations; shifting their behaviour more frequently at the earliest stages of their community life and converging more towards the linguistic styles of others, until they gradually stabilise and become more attuned to the community norms, perhaps even progressing to leadership roles themselves. A useful future investigation would be to track the progress of the Zelig-like Non-Leaders over time, to see if their propensity to adapt and change their linguistic style affects their ability to progress within the community, e.g. does linguistic style convergence enable them to earn reputation more quickly, or provide an indicator of longevity within the community?

5 Conclusions and Future Work

"...and it shows exactly what you can do if you're a total psychotic."
- Leonard Zelig, Zelig (Allen, 1983)

We have presented a metric for measuring linguistic accommodation in a systematic manner, considering not only the context of an individual pair's communications, but the background models for both individuals. Thus, the Zelig Quotient provides an objective, quantifiable measure of convergence and divergence in language use between individuals, as defined by the movement in an individual's typical linguistic style towards or away from the typical linguistic style of their conversational partner. The Zelig Quotient is meaningful for differentiating between those who typically accommodate their language use towards many people (Zelig-like individuals) from those who don't. In addition, the metric includes calculation of pair-wise author to reader accommodation scores. Thus, a full picture of how an individual is behaving in terms of convergence and divergence can be gained by examining these pair-wise scores. In combination, these two scores together provide a comprehensive and illuminating picture of an individual's accommodation behaviour. This provides a framework for investigating the circumstances surrounding such variation in language use, over large datasets, in a manner which has not previously been undertaken.

We acknowledge that community forums such as these may not be the ideal dataset for evaluating dyadic accommodation, as authors may be addressing multiple people. However, it is worth noting the use of the Zelig metric seems to be effective even in this community forum dataset. Work on testing the metric on a wide range of existing datasets, including courtroom interactions and dyadic therapeutic interactions, is ongoing.

There are a number of additional possibilities for future work. Firstly, although in the current study we focussed on linguistic features that are stylistic in nature, it would be simple to alter the Zelig metric to use greater, fewer or entirely different linguistic features, for instance to explore semantic or content aspects of language. We have so far considered only a small set of linguistic style features, and it may be worth expanding this to a greater variety; in particular, concentrating on features which have been shown to be the subject of accommodation in sociolinguistic studies (such as Bunz & Campbell, 2004). We also intend to test the metric with languages other than English in future research. Secondly, the separation between our two sample groups in terms of Zelig distributions suggests that this Quotient may be useful as a predictor of group membership/reputation score. A considerable body of work has attempted to accurately classify community members based on their communication behaviours and our results suggest that Zelig may be useful in this domain. We also intend to investigate in more detail what kinds of relationships are characterized by higher levels of accommodation, to see whether this accords with underpinning theories of politeness and social identity. Group dynamics, including linguistic accommodation to group norms in multiparty communication, and the individual's contribution to constructing a group identity, is also a large area ripe for further study.

References

Erik Aumayr, Jeffrey Chan, and Conor Hayes. 2011. Reconstruction of threaded conversations in online discussion forums. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-11)*, pages 26–33.

Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. 2002. Language trees and zipping. *Physical Review Letters*, 88(4):1 – 4.

James Bradac and Howard Giles. 2005. Language and social psychology: Conceptual niceties, complexities, curiosities, monstrosities, and how it all works. In K. L. Fitch and R. E. Sanders, editors, *Handbook of language and social interaction*, pages 201 – 230. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Joseph N Cappella. 1979. Talk-silence sequences in informal conversations i. *Human Communication Research*, 6(1):3–17.

Kenneth W Church. 2000. Empirical estimates of adaptation: The chance of two noriegas is closer to p/2 than p2. In *Proceedings of the 18th Conference on Computational Linguistics (COLING2000)*, volume 1, pages 180 – 186.

Noah Coccaro and Daniel Jurafsky. 1998. Towards better integration of semantic predictors in statistical language modeling. In *Proceedings of International Conference on Spoken Language Processing (ICSLP-98)*, pages 2403 – 2406.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76 – 87. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words!: Linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754.

Amit Dubey, Patrick Sturt, and Frank Keller. 2005. Parallelism in coordination as an instance of syntactic priming: evidence from corpus-based modeling. In *Proceedings of the Human Language Technology conference and the conference on Empirical Methods in Natural Language Processing*, pages 827 – 834.

Simon Garrod and Martin. J Pickering. 2004. Why is conversation so easy? *Trends in Cognitive Sciences*, 8(1):8 – 11.

Jessica Gasiorek and Howard Giles. 2012. Effects of inferred motive on evaluations of nonaccommodative communication. *Human Communication Research*, 38(3):309 – 331.

Howard Giles and Tania Ogay. 2007. Communication accommodation theory. In Bryan B Whaley and Wendy Samter, editors, *Explaining communication: Contemporary theories and exemplars*, pages 325 – 345. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. In Howard Giles, Justine Coupland, and Nikolas Coupland, editors, *Contexts of accommodation: Developments in applied sociolinguistics*, pages 1 – 68. New York: Cambridge University Press.

Howard Giles. 1973. Accent mobility: A model and some data. *Anthropological linguistics*, 15(2):87–105.

Augusto Gnisci. 2005. Sequential strategies of accommodation: A new method in courtroom. *British Journal of Social Psychology*, 44(4):621 – 643.

David Huffaker, Joseph Jorgensen, Francisco Iacobelli, Paul Tepper, and Justine Cassell. 2006. Computational measures for language similarity across time in online communities. In *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech*, pages 15–22. Association for Computational Linguistics.

Molly E Ireland, Richard B Slatcher, Paul W Eastwick, Lauren E Scissors, Eli J Finkel, and James W Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1):39–44.

Elizabeth Jones, Cynthia Gallois, Victor Callan, and Michelle Barker. 1999. Strategies of accommodation: Development of a coding system for conversational interaction. *Journal of Language and Social Psychology*, 18(2):123 – 152.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97 – 133.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444.

Michael Natale. 1975. Convergence of mean vocal intensity in dyadic communication as function of social desirability. *Journal of Personality and Social Psychology*, 32(5):790–804.

Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*.

David Reitter, Frank Keller, and Johanna Moore. 2006. Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology conference of the North American chapter of the Association for Computational Linguistics*, pages 121– 124.

Lauren E Scissors, Alastair J Gill, and Darren Gergle. 2008. Linguistic mimicry and trust in text-based cmc. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 277–280. ACM.

Svetlana Stenchikova and Amanda Stent. 2007. Measuring adaptation between dialogs. In *Proceedings of the 8th SIGdial workshop on Discourse and Dialogue*, pages 166 – 173.

Richard L Street. 1984. Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research*, 11(2):139–169.

Richard L Street. 1991. Accommodation in medical consultations. In Howard Giles, Justine Coupland, and Nikolas Coupland, editors, *Contexts of accommodation: Developments in applied sociolinguistics*, pages 131 – 156. New York: Cambridge University Press.

Donald M Taylor, Lise M Simard, and Danielle Papineau. 1978. Perceptions of cultural differences and language use: A field study in a bilingual environment. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 10(3):181.

Marinus Van den Berg. 1986. Language planning and language use in taiwan: social identity, language accommodation, and language choice behavior. *International journal of the sociology of language*, (59):97–116.

Michael Willemyns, Cynthia Gallois, Victor Callan, and J Pittam. 1997. Accent accommodation in the employment interview. *Journal of Language and Social Psychology*, 15(1):3 – 22.