Computational Linguistics in the Netherlands Journal 1 (2011) 95-108 Submitted 7/2011; Published 12/2011

Im chattin :-) u wanna NLP it: Analyzing Reduction in Chat

Hans van Halteren HVH@LET.RU.NL
Craig H. Martell CMARTELL@NPS.EDU
Du Caixia SHY2005MP@163.COM
Yan Gu ENJOYSTUDYING@HOTMAIL.COM
Johan Kobben J.KOBBEN@GMAIL.COM
Leequisach Panjaitan LEEQUISACH.AVEN.PANJAITAN@GMAIL.COM
Louise Schubotz LOUISE.SCHUBOTZ@MPI.NL
Kateryna Vasylenko K.M.VASYLENKO@GMAIL.COM
Yuliya Vladimirova YULIA.U.VLADIMIROVA@GMAIL.COM
Centre for Language Studies
Radboud University Nijmegen

Abstract

In this paper, we[1] study the conditions in which words are reduced in chat text, in particular whether these conditions resemble those found for reduction in spoken text. We extract instances of five types of reduction from the 2Mw NPS Chat corpus, and apply regression analysis to measure the level of influence for 9 potentially predicting features. Although we do find significant effects, there appears to be no relation to the findings for speech. Furthermore, only the feature representing individual preferences is shown to have a substantial explanatory value. We conclude that for a reliable prediction on a chatter's choice to use reduction, we need to look further than our limited set of features and apply more intricate analyses, both of which necessitate access to more extensive and more accessible chat data.

1. Introduction

In traditional linguistics, and also in traditional natural language processing (NLP), we generally see a distinction between written and spoken language. There are a few intersections, such as scripted speech or dialogue in fiction, but in principle these are written language, possibly rendered later in spoken form (even though the author may well have attempted to imitate properties of spoken language use). A related aspect is that of preparedness. Mostly, written language is constructed with more forethought and possibly over various stages of editing, while real spoken language tends to be more spontaneous. Recently, however, new types of communication are becoming available which do not conform to these traditional notions. We will focus on the clearest example, chatroom conversations. The nature of the text produced there is clearly very similar to spoken dialogue, being a real time multi-participant discussion where utterances are spontaneously produced under time pressure (Blakeman 2004). On the other hand, the mode of communication is not speech,

1. This study was done during the course Corpus Based Methods in the Research Master Language and Communication at the Radboud University Nijmegen and the University of Tilburg. Apart from van Halteren (Radboud University; hvh@let.ru.nl) and Martell (Naval Postgraduate School, Monterey, CA), all authors were students in this course.

but typed text. Chat text allows us to study whether phenomena observed in spontaneous spoken language use also occur when produced by typing rather than speaking. A better understanding of this will also be of great use for language technology, e.g. in building automatic agents that are expected to process and produce natural text in conversations with humans.

One of the more apparent phenomena in spoken text is that speakers exhibit variation in their speech, often in the form of more or less reduced forms of the uttered words. In normal written text, such reduction is not present, but words that are spelled using fewer characters than the conventional spelling are observed frequently in chatroom conversation. It is not uncommon to find sentences like the following:

1. *u should've told me 15 mins b4* ("you should have told me 15 minutes before")

2. *wots ur country?* ("what is your country?")

The link to speech is clear, as the reduction makes use of the phonological characteristics of the words. The main question we want to pursue in this paper, by way of a corpus study, is whether the conditions under which reduction (defined as spelling a word with fewer characters than the 'normal', accepted spelling) occurs are the same as those observed for spoken language. In other words: Are factors which influence the level of reduction in speech also active in chat? Additionally, we want to get an impression of the degree to which the use of reduction can be predicted, for the already mentioned purposes of dealing with reduction in NLP. Therefore, we will also include some factors not based on the literature on speech, but on other sources, including our own intuitions.

The remainder of this paper is structured as follows. In Section 2 we formulate some hypotheses about reduction in chat which ought to be correct if the factors in speech transfer to chat. In Section 3 we describe the corpus material we use for our examinations and initial findings from a manual inspection of part of the corpus. In Section 4, we present the selection of five types of reductions we will subject to statistical analysis and the composition of the features we take into account. In Section 5 we present the outcome of a regression analysis and in Section 6 discuss this outcome in the light of our hypotheses. In Section 7, finally, we list our main conclusions.

2. Reduction in Chat Language

In our investigation into reduction in chat language, we first put forward a number of hypotheses on factors which we expect to influence the level of reduction. As explained above, our main source of these factors is the literature on reduction in speech. However, this literature is far too extensive to discuss here in full. We will therefore organize this section around the hypotheses and mention only (examples of) literature that has a bearing on the hypothesis in question. For more information on reduction in speech, we advise Shockley (2003) and Ernestus (In Press) as good starting points. For some hypotheses, speech does not provide sufficient clues, but we can base them on other literature or our intuitions. Before we list our hypotheses, note that these hypotheses refer to the choice between reduced and full forms when both forms are still regularly used, i.e. when there is

an alternation, and do not refer to reduced forms which have become part of the standard chat vocabulary, such as *lol* ("laughing out loud").

Our first hypothesis is that ***(1) the habits of the author will be a main factor in the level of reduction***. Just as for speech, there is nothing in the linguistic aspects of a message or in the situational ones (other than in SMS or tweet) that really forces an author to reduce. Personal behavioral patterns will be of most influence on when and where the author reduces. It may well be, however, that we can find trends for similar reduction behavior in groups of authors with specific common properties.

A trend that comes to mind quickly is that ***(2) younger people will reduce more than older people***. This has been observed in speech, e.g. by Berglund (2000), who found that younger speakers are more likely to use the (reduced) non-standard form *gonna* rather than *going to* than older speakers, and Bell et al. (2003), who found the same tendency as a general influence on reduction. For chat, this is also intuitively likely as the younger chatters can be assumed to have had more contact with new media such as SMS and have a less fixed spelling pattern. As for the other main characteristic of an author, gender, formulating a hypothesis is harder. For speech we see potentially contradictory claims. Bell et al. (2003) observe that men are generally more likely to use reduced forms than women, but Ernestus (In Press) states that women use reduced variants more often when they are more prestigious or considered to be the norm. Seeing that it is hard to determine which variants are prestigious in chat, we refrain from formulating a firm hypothesis and restrict ourselves to ***(3) gender may have an influence on the level of reduction***. Another characteristic that may be of importance in our specific investigation is the frequency with which an author is participating in chats. We have discovered no previous information, but our intuition would be that ***(4) more frequent chatters will reduce more than less frequent chatters***.

Going from the personal to the situational (and hence staying within the extralinguistic influences), we observe that the chatters participating in a particular discussion can be expected to conform to a common behavior. First of all, they need to take care that the other participants will actually understand the reduced forms they are producing (Grice 1975). But also, and especially so for some topics, they will try to do their best to appear to belong to the group, in this case by conforming to peer behavior and producing the same type and level of reduction as found in the chat. In terms of a hypothesis, this means ***(5) the use of reduced forms will be patterned on previous uses of reduced forms in the same chat***.

Moving on to the linguistic factors, we first predict that ***(6) more predictable words are more likely to be reduced than less predictable words***. This is again a well-known factor in speech reduction (Bell et al. 2009). Basically, the more common an item is, the more likely it is to reduce if it contains elements which are reduction-prone. The frequency dependent reduced form has for instance been observed in a large digitized corpus of American English by Greenberg and Fosler-Lussier (2000), who link it to the observation that the brain appears to process words of high frequency more quickly than their infrequent counterparts and hypothesize that therefore frequent words may need to be less fully specified in order to achieve adequate communication. This might be one of the reasons why *walking* is often reduced as *walkin* or *and* as *n*, as opposed to infrequent words such as *equivalent* or *accomplish*. This presumption is also supported by Bell et al. (2009). Bell et

al. also observe that more reduction is seen in function words than in content words, and that the predictability affecting reduction is not limited to the word itself. They show that the immediate context has an influence on the reduction of words: predictability on the basis of the following words determines the reduction of both content and function words, but only very frequent function words are sensitive to predictability on the basis of the preceding word. However, words other than the immediate neighbors of the word in question do not seem to have an influence on reduction.

We must also take into account that the potentially reduced word is embedded in chat. First of all, chat has been developing its own vocabulary and for many words, there are clear biases towards either full or reduced form, so that ***(7) for each (type of) word in question there will be a very influential bias towards or away from reduction***. Furthermore, the overall placement of the potentially reduced word in a chatroom discussion may also be of influence. If chat is similar to speech, we should expect that ***(8) there will be more reduction earlier in a post than later in a post***, as this position effect has been noted for speech (Shockley 2003). On the other hand, we should note that Bell et al. (2003) found that position within an utterance also had another type of influence on realization of a word: words that occur in utterance initial or final position are less likely to be reduced than words that occur in medial position.[2] For chat itself, is has been suggested (Shockley 2003) that as the aim of reduction is the speeding up of the typing process, one could expect that ***(9) longer posts will show more reduction than shorter posts***.

There are yet many other factors that might be of influence. In speech, given information is more often reduced that new information (Chafe 1987). Some topics might attract more reduction, whether or not by way of the nature of the chatters participating in the chat (Shockley 2003). The urgency felt in finishing a particular post in a particular discussion might encourage a chatter to reduce more. Particular syntactic properties might be in play (Lakoff 1970, Warren et al. 2003), or others such as stress, timing, syllable structure and higher level discourse effects (Shockley 2003). And so forth. However, in the current investigation, we have to ignore these factors as we have no means to identify them for the specific posts in our corpus study (see the following section).

3. Corpus Material and Method

In order to investigate whether our hypotheses hold in chat, we used the NPS Chat Corpus of North American English chat conversations (Forsyth and Martell 2007). The corpus was collected at the Naval Postgraduate School in Monterey, CA in 2006. In total, this corpus consists of approximately 500,000 chat posts gathered from various online services. However, only part of this corpus has been processed: privacy masked, PoS tagged, as well as dialogue-act tagged. This part, 10,567 posts from age specific chat rooms, forms Release 1.0 and has been made publicly available via the website http://faculty.nps.edu/cmartell/NPSChat.htm. Future releases are to contain more posts from more domains.

To get a first impression of the frequency and type of reductions which can be found in a corpus like this, we manually annotated approximately 8,000 words from 2132 posts of user generated text, about 500 posts each from the chatrooms `talkcity-20s`, `talkcity-adults`,

2. It should be pointed out that the word 'utterance' has a different interpretation for chat (post) than for speech (usually turn in dialogue and sentence in monologue).

`14-19teens` and `40splus`. Using a limited preliminary tag set, we marked all non-canonical word forms. At this point, we were not only focusing on reductions but still also considering other spelling variation present in the data.

Among the approximately 8000 words from the corpus which we inspected, about one in eight words (ca. 970) were non-standard forms (see Table 1).

Table 1: Non-standard forms found in the corpus.

Type	Examples	No. of instances
Chat specific abbreviations	*lol, brb, lmao, wb*	232
Emoticons	*;-), :), :-@, *blush*, <hugs>*	68
Exclamations	*haha, awww, geeesh, ewww*	104
Spelling variants	*u, wanna, im, bout, cant, donno, gurls, dawg*	500
Unclassifiable	*kawing, chocha, cuddlicious*	67
		971

One rather large group is formed by what could be called 'chat specific abbreviations', i.e. mostly acronyms of short phrases which are frequently used in chat and which are highly conventionalized. Examples of these items are e.g. *lol* standing for "laughing out loud", *pm* for "private message", *brb* for "be right back", or *wb* for "welcome back". Since these forms are so conventionalized and practically completely replace their full forms, they do not constitute the kind of reduction we were interested in.

Another group of non-standard items were so-called emoticons: little smiley faces like for example *:-)*, *;)*, or *:-@*, each expressing a different emotion, as well as other expressions of emotion, e.g. **blush**. Like the chat specific abbreviations, these items are highly conventionalized and are not creatively composed on the fly to express the users' emotions.

The next group of non-canonical forms we encountered and termed Exclamations showed more individual variation. It contains forms like *haha* of various lengths, *awww*, or *geesh*. Again, these forms are not reductions and were therefore not considered for further analysis.

By far the largest group of non-standard items, with approximately 500 instances, were spelling variations of existing words, partly reductions but also including other variations. Word forms spelled deviantly but non-reduced were for example *gurls* for "girls", *nope* for "no", or *dawg* for "dog".

As to the reduced spellings, most have a relation to speech. Most common are the spellings reflecting the colloquial pronunciation of certain words or phrases, such as e.g. *wanna* for "want to", *gonna* for "going to", *talkin* for "talking", or *ya* for "you", all of which appear to be lexicalized. Then we find quite a lot of phonetically inspired creative (but sometimes also already lexicalized) spellings, such as e.g. *nite* for "night", *u* for "you", *r* for "are" and forms in which part of a word is replaced by a digit, such as e.g. *2DAY* for "today". However, there are also reductions that are based purely on writing. The most common example here is the dropping of the apostrophe in forms like *cant* instead of "can't" and *im* instead of "I'm". All of these reduced forms alternate with their full forms and can therefore be used to study the factors that influence the choice between full and reduced form.

4. Cases for Statistical Analysis

Obviously, the number of occurrences in the inspected part of the corpus is much too low for a sensible statistical analysis. Fortunately, although the total collection of 500,000 chat posts in possession of the Naval Postgraduate School is not open to public inspection for reasons of protection of privacy, we were kindly allowed to extract all instances of specific cases by means of an automatic script, along with a number of factors we considered to be potentially influential for the reduction.

Some care had to go into the selection of these specific cases, as it was clear that we would be able to run the automatic script only on a very limited number of times. In order to control for the type of reduction, we have to select specific classes within which the cases can be expected to behave similarly. We formulated three criteria for our final selection of word forms to be analyzed statistically: 1) they had to be likely to occur in sufficient numbers, to allow for statistical modeling in the first place, 2) they should be used by more than five percent of the users, to prevent very frequent use by only a non-representative group of users, and 3) it should be possible to extract them automatically, as we had neither access nor the means to inspect the 500,000 chat posts manually. In line with these criteria, we chose five cases for analysis, listed in Table 2.

Table 2: Cases selected for analysis. Numbers given for the manually inspected part of the corpus.

Type	Examples	No. of occurrences	No. of different users
TO	*wanna, gonna*	37	25
YA	-	10	7
U	-	37	15
APOS (apostrophe drop)	*im, dont, didnt, whats*	56	25
ING (g drop in present participle)	*talkin, goin, wantin, havin*	33	19

Similarly, features had to be selected which could be used to attempt to predict reduction. Here the criteria were 1) the feature is expected to be potentially influencing reduction, as indicated by our hypotheses above, 2) the feature can be extracted automatically from the corpus, and 3) the feature does not compromise the privacy restrictions which have been placed on the full corpus. The last criterion prevented us from extracting the full context of each potential reduction and we had to compromise on a mere two words to the left and right of the potentially reduced word.

The final selection of features is listed in Table 3.

Table 3: Selected features potentially influencing reduction

Pbias	General bias for reduction by this poster, calculated over all reduction cases for the type in question for this author.[3]
Fbias	General bias for reduction for this word form, calculated over all reduction cases for this form.
Age	Age of author.
Gen	Gender of author, with male (arbitrarily) represented by 1 and female by 0.
Rprev	Previous choice for same alternation in same chatroom.[4]
Len	Length of post.
Pos	Position in post, represented by length of full left context divided by length of full right context.
Lprob/Rprob	Probability of full form given the immediate two-word left/right context, based on N-gram counts in the Google Web 1T N-gram corpus. When trigrams are unavailable, we use a fallback strategy to bigrams and unigrams.[5]

Once the set of automatically extracted cases was available, we double-checked (semi-automatically) whether they indeed constituted instances of the types of reduction which we had been looking for. The resulting numbers of cases are shown in Table 4.

Table 4: Number of cases collected for each investigated reduction type.

Type	Number of cases after filtering	Number of cases with full form	Number of cases with reduced form
To	10322	3886 (38%)	6436 (62%)
Ya	34960	31628 (90%)	3332 (10%)
U	40601	31628 (78%)	8973 (22%)
Apos	45845	23410 (51%)	22435 (49%)
Ing	30224	25445 (84%)	4779 (16%)

5. Regression Analysis

For the cases listed in Table 4, we attempted to determine whether the collected features were of influence on the choice between reduced and full form. We did this separately for each of the five case types (to, ya, u, apos, and ing). Each time we built three types of regression model.[6] First we built logistic regression models predicting the choice per case on the basis of each individual feature, then we built a logistic regression model for each

3. Pbias often represents a rather rough approximation as the number of cases on which it is based may be very low.
4. The extracted data does not include markers for beginning and end of individual chats, so that the previous choice may have been in another chat.
5. This implementation does not match the notion of contextual predictability as present in the literature on speech completely. As for the context window, we opt to use as much as available so that we could test whether the limitations observed in speech are present in chat too. As for the calculation of the probability, we had to base ourselves on the available data; more on this in Section 6.
6. We used R (R Development Core Team 2008), more specifically the functions `lrm` and `lm`.

case but with a combination of features using forward selection, and finally we built a linear regression model at the author level, predicting Pbias on the basis of author characteristics.

In order to get a first indication of which features might play a role, we built models using individual features as single predictors for each kind of reduction we investigated. The results for the various models are shown in Table 5.

Table 5: Results for logistic regression modelling of reduction on the basis of each feature individually. In the table the + or - indicates whether the influence of the factors on the level of reduction is positive or negative. ·,*, ** and *** stand for significance level at p =0.1, p =0.05, p =0.01 and p =0.001 respectively. n.s. stands for not significant and n.a. means not applicable, as the type of reduction in question does not have varying forms. R^2 is the squared correlation coefficient, which quantifies (on a scale of 0 to 1) the proportion of the variance that the model explains.

Type	Pbias	Fbias	Gen	Age	Rprev
TO	+ *** R^2 =.685	+ *** R^2 =.004	n.s.	- ** R^2 =.005	+ *** R^2 =.016
YA	+ *** R^2 =.370	n.a.	n.s.	+ *** R^2 =.008	+ *** R^2 =.023
U	+ *** R^2 =.795	n.a.	n.s	- *** R^2 =.012	+ *** R^2 =.053
APOS	+ *** R^2 =.842	+ *** R^2 =.086	+ *** R^2 =.014	- *** R^2 =.049	+ *** R^2 =.046
ING	+ *** R^2 =.484	+ *** R^2 =.154	+ *** R^2 =.046	- *** R^2 =.003	+ *** R^2 =.021

Type	Len	Pos	Lprob	Rprob
TO	- *** R^2 =.020	+ ** R^2 =.001	- *** R^2 =.027	+ *** R^2 =.005
YA	- *** R^2 =.014	+ *** R^2 =.025	n.s.	- *** R^2 =.030
U	- *** R^2 =.031	+ *** R^2 =.001	- *** R^2 =.005	- *** R^2 =.008
APOS	- *** R^2 =.003	n.s.	+ *** R^2 =.003	+ *** R^2 =.002
ING	- *** R^2 =.027	n.s.	- *** R^2 =.002	- *** R^2 =.008

As can be seen in the table above, the biases are of major importance. The Pbias is very significant in all cases and, furthermore, with an R^2 which is very high for models based on a single feature. This indicates that the amount of reduction depends heavily on the individual chatter. The second factor, Fbias, is also shown significant for all cases where it is applicable. Gender only shows significance for APOS and ING, and with very weak explanatory value. Age, however, is always significant, with the expected negative effect, except in case of YA, where we find a significant positive effect. The previous choice in the chatroom also shows a significant positive effect for all types. It must be said, however, that for all these three predictors, the explained variance is rather low. Progressing to the linguistic features, the post length is found to have a significant negative effect, indicating

that the longer the post, the lower the likelihood of reduction, contrary to our expectations. Our expectations are also not fulfilled for the position in the post as there is either no effect or even a positive one, for TO, YA and U, which means that there is more reduction when going closer to the end of the post. As for Lprob or Rprob, we cannot say we see any clear pattern emerge. Again, it should be noted that for these as well as for the other linguistic factors, significance levels may be high, but explanatory value very low.

In the models above, we used each individual feature by itself. Under these circumstances, we have to be careful that any observed regularities are not caused by correlation with other features, which in our case may well be so given the very low explained variance for quite a few models. We therefore decided to also build a combined multiple regression model. For each type of reduction, we built a forward model: we started with the most significant individual feature and repeatedly added that feature which improved the model the most, stopping when no further improvement was observed. The significance for the features in the resulting combined models is shown in Table 6.

Table 6: Results for logistic regression modelling of reduction using a combined model, built with forward selection of features.

Type	Pbias	Fbias	Rprev	Len	Pos	Lprob	Rprob	R^2
TO	+ ***	- *	n.s.	- **	n.s.	- ***	+ ***	.703
YA	+ ***	na	+ **	- ***	+ ***	n.s.	- ***	.419
U	+ ***	na	+ ***	- ***	n.s.	n.s.	n.s.	.804
APOS	+ ***	+ ***	n.s.	- ***	n.s.	n.s.	n.s.	.872
ING	+ ***	+ ***	+ ***	- **	n.s.	- *	n.s.	.587

As we see, for some features which appeared significant when considered by themselves, we cannot show significance anymore when they are combined with other features. For age and gender, this could be expected, since their influence is clearly embedded in Pbias. For various cells for Rprev, Len, Pos, Lprob and Rprob, an explanation is harder to find. Possibly the original significance was indeed due to some chance correlation, but it is also possible that the individual influence was so weak that it is now overwhelmed by that of the stronger features and can just not be observed any longer. It is well-known that combined models are often hard to interpret (Rietveld and van Hout 2005) and for now, we do not want to rule out that observed individual influences have some real existence, although obviously very weak. The complex nature of combined models must also be called upon to explain the fact that two significant effects change polarity in the combined models in relation to the individual models: Fbias now shows a negative effect for TO and Rprob does the same for ING. For the probabilities we already stated that we could discern no pattern and the polarity change only confirms this further. What strange interference could explain the polarity change for the Fbias cannot be found out without more detailed statistical methods, which are beyond the scope of this paper.

As already stated, it is not surprising that age and gender lose any significance they have when included in a combined model. After all, they are characteristics of the author and as such will already be covered in the feature Pbias. To see to what degree they play a role as a component of the author's behaviour, we attempt to predict Pbias on the basis

of age and gender, as well as the author's activity level. We measure the latter in terms of the log of the number of cases we have for each specific author, assuming that the number of posts in which one of our target alternations is present is a good approximation of an author's total number of posts. For this model, we restrict ourselves to those authors for whom we know the age and gender and who have at least 10 cases in our experimental data. As a result, the number of cases on which we are building our models is substantially lower than the number of cases in the models above.

Table 7: Results for logistic regression modelling of Pbias on the basis of author characteristics, using a combined model with all features.

Type	Number of cases	Age	Gender	Activ	R^2
TO	134	n.s.	n.s.	- ***	.101
YA	379	+ *	n.s.	n.s.	.022
U	398	n.s.	n.s.	- *	.020
APOS	315	- ***	n.s.	- *	.068
ING	397	- ·	n.s.	n.s.	.014

Just like for the models at case level, age is shown to have significant influence. However, the significance level is lower, and significance can even no longer be shown for TO and U, but this might be explained by the much smaller number of data points, which tends to make it harder to prove significance. Gender shows no significance at all, but the activity level does appear to play a role. It shows a negative effect for especially TO, but also for U and APOS. Note that a negative effect means that, contrary to our intuition, more active users actually reduce less.

6. Discussion

As so often, statistical analysis does not provide all definitive answers, but also tends to throw up some further questions and even outright confusion. Still, we have gained new information which we can compare with our hypotheses. We again organize our discussion according to the hypotheses from Section 2.

(1) The habits of the author will be a main factor in the level of reduction. Our first hypothesis is also the one that is most unequivocally confirmed. Across all types of reduction, the feature Pbias was both highly significant and showing high explained variance.

(2) Younger people will reduce more than older people. As for age (as well as for gender), we got input from two different statistical models: that at case level and that at author level. The author level does not always confirm what we see at case level, but this might be due to the low number of cases used to build the author models. Our hypothesis about age has been partly confirmed. Younger people do reduce more than older people, with highly significant effects present for TO, U, APOS and ING at case level, but only for APOS at poster level. Note that we are not claiming that people start reducing less as they get older. The only conclusion we can draw that at this time older people are reducing less; the now younger might well retain their reduction behavior over the years and keep

reducing as they do now. However, we also found in both analyses, in contradiction with the hypothesis, that *ya* is used less by younger people. A reason for this might be that *ya* is only used by older posters and is perceived by younger posters as old-fashioned, but we would need to go back to the primary data to check where and by whom *ya* is used in order to discover whether this is the real and only reason.

(3) Gender may have an influence on the level of reduction. For gender we did not have a clear hypothesis, and that decision appears to have been wise. Although men appear to reduce more with APOS and ING when we look at the case level, this finding is unconfirmed at poster level. If any effect is present, it might well be indirect, e.g. by way of the topics that are discussed by men and women. Here too, detailed examinations of primary data are needed.

(4) More frequent chatters will reduce more than less frequent chatters. A closer look at the data is also desired for the influence of posting activity, since our hypothesis has been not only unsupported, but even contradicted. The analysis at the poster level shows that more active users appear to reduce less, with significance at the 0.001 level for TO, and 0.05 level for U and APOS. Seeing how wrong our intuition can be where it comes to chatting behaviour, we do not venture any explanations without a further examination of actual chats.

(5) The use of reduced forms will be patterned on previous uses of reduced forms in the same chat. When examining previous reduction behaviour in the chatroom by itself, the hypothesis appears correct, with highly significant positive effects for all types. However, Rprev loses its significance for TO and APOS when Pbias is also in the model. A first suggested explanation for this, namely that Rprev just mirrors the specific poster who is of course in the same room, does not appear to be the only one. For when we repeated the analysis with the definition of Rprev changed to being the previous choice in same room by another poster, we found that significance levels generally stayed the same. Here we would suggest more fine-grained modelling techniques as a next step.

(6) More predictable words are more likely to be reduced than less predictable words. Our hypothesis that more reduction will be found in more predictable words was neither confirmed not contradicted. In general, the results of the regression analysis were highly ambiguous. In hindsight, we put forward that a probability calculation based on the Google N-Gram data was probably not suitable for our purposes. Not only is chat text very different from average web text, but the N-gram data also contains little to no predicting bigrams or unigrams which are themselves reduced, leading to complications in the probability estimate. Note, however, that a better hypothesis based on the full chat corpus itself was outside our reach as the full corpus is not available.

(7) For each (type of) word in question there will be a very influential bias towards or away from reduction. We have indeed found support for our hypothesis, at least for the two reduction types where it is most relevant: for both APOS and ING, both the individual and combined model show high significance of this feature. For TO, the situation is less clear, with a significant positive effect in the individual model and a negative one in the combined model, but it should be noted that there are only two forms (*wanna* and *gonna*) and there are probably syntactic factors in play (as mentioned above) which we do not account for in our models. And for U and YA, there is no Fbias, since there is only one form present.

(8) There will be more reduction earlier in a post than later in a post. Contrary to our hypothesis, the regression models showed no decrease of reduction as we progress along a post. For some types, the models even showed an increase, at the 0.05 significance level. Our hypothesis was based on spoken text, and was clearly linked to the topic-comment structure of most utterances, combined with the observation that given material is more easily reduced. Having seen more chat text now, we are not surprised that our hypothesis does not hold, as this structure is much less present in chat posts (cf. also Ye 2005). We do not yet propose an explanation for the observed increase but again suggest examining the text data more directly than by way of a regression model.

(9) Longer posts will show more reduction than shorter posts. This final hypothesis was also not only not confirmed but even contradicted: longer posts show less reduction than shorter posts, with high to very high significance for all types. It would appear that, although there may well be more reduction in longer posts in absolute terms, the relative amount of reduction is in fact lower. We expect that an explanation may be found in the nature of the chats in question: successions of quick fire short (and highly reduced) posts in chats of a more social nature with low content versus extensive argumentation using longer posts in chats with higher content and exhibiting a structure more like 'normal' written language. However, again, we would have to inspect the primary data to confirm this expectation, after using more detailed statistical methods to identify the posts most likely to provide relevant information.

7. Conclusion

In this paper, we set out to find some regularities in the use of reduced word forms in chat language. We were especially interested whether these regularities resembled those observed for reduction in speech, chat and speech both being spontaneous and real time forms of language production. Furthermore, we wondered whether the use of reduction could be predicted to a sufficient degree that the predictions could help process and generate chat text, e.g. in the context of automated systems such as chatbots.

Using regression modeling, we did find several regularities for the use of reducing "want to" and "going to" to *wanna* and *gonna*, expressing "you" as *u* or *ya*, dropping the *g* in present participles and dropping the apostrophe in pronoun-verb combinations. The most important factor proves to be the individual author's preferences, which themselves appear to be partly influenced by age (younger reducing more than older, with the exception of *ya*) and the amount of chatting done (more active reducing less than more active). We have not been able to show an influence of gender. The word forms in question also show clearly influential biases, as does their placement in posts: longer posts show relatively less reduction and the degree of reduction increases as the post progresses. Unlike for speech, we have not found a clear influence of the predictability of the word form on the basis of its context.

In general, various hypotheses that we made on the basis of what is known for speech did not hold for chat. In some cases, we even found opposite effects. For those hypotheses that did (mostly) hold, relating to authors, word forms and age, it is very well possible to suggest explanations that that do not rely on a comparison to speech. It would seem that, although the same basic reasons for reducing may be active and reductions do seem

to borrow from the phonological structure of the words, the circumstances under which reduction is used in chat are certainly not the same as those in speech. The view of chat as merely keyboarded speech is clearly an oversimplification as chat has apparently developed into a separate text type with an idiosyncratic nature of its own.

As for the predictability of reduced forms in chat, without reference to speech, there is clearly still quite some work to do. The author's preferences account for a large to very large part of the variance (cf. Table 5), but its examined component factors hardly seem to contribute a substantial influence (cf. Table 7), so that we can only draw the general and vague conclusion that idiolect is important. The factors outside the idiolect hardly help improve on the explained variance (cf. Table 6). We tend to conclude that the view we have on the data is much too narrow. Further factors are likely to be found in the full context, both inside the current post and the posts preceding it. To find these factors, however, we need to be able to understand such aspects as the intended message, its context, and the interplay between the participants in the chat. For this, we will need full access to chat corpora, be it the NPS Corpus or preferably even larger ones. And as privacy considerations will keep playing a role in text types like this for the foreseeable future, we suggest a serious investment in both convincing contributing chatters to allow full access and developing adequate automatic anonymization tools (similar to what is being done in medical research, cf. Meystre et al. 2010).

References

Bell, A., D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea (2003), Effects of disfluencies, predictability, and utterance position on word form variation in english conversation, *Journal of the Acoustical Society of America* **113**, pp. 1001–1024.

Bell, A., J.M. Brenier, M. Gregory, C. Girand, and D. Jurafsky (2009), Predictability effects on durations of content and function words in conversational english, *Journal of Memory and Language* **60**, pp. 92–111.

Berglund, Y. (2000), *Gonna* and *going to* in the spoken component of the British National Corpus, *in* Mair, C. and M. Hundt, editors, *Corpus linguistic and linguistic theory: papers from the twentieth international conference on English language research on computerized corpora (ICAME20), Freiburg im Breisgau, 1999*, Rodopi, Amsterdam/Atlanta, pp. 35–49.

Blakeman, A. (2004), *An investigation of the language of Internet chat rooms*, PhD thesis, Lancaster University.

Chafe, W. (1987), Cognitive constraints on information flow, *in* Tomlin, R.S., editor, *Coherence and Grounding in Discourse*, John Benjamins, Amsterdam/Philadelphia, pp. 21–51.

Ernestus, M. (In Press), Acoustic reduction and the roles of abstractions and exemplars in speech processing, *Lingua*.

Forsyth, Eric. M. and Craig H. Martell (2007), Lexical and discourse analysis of online chat dialog, *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC) 2007*, pp. 19–26.

Greenberg, S. and E. Fosler-Lussier (2000), The uninvited guest: Information's role in guiding the production of spontaneous speech, *Proceedings of the Crest Workshop on Models of Speech Production: Motor Planning and Articulatory Modeling, Kloster Seeon, Germany, May 1–4, 2000*.

Grice, H.P. (1975), Logic and conversation, *in* Cole and Morgan, editors, *Syntax and Semantics 3: Speech Acts*, Academic Press, New York.

Lakoff, G. (1970), Global rules, *Language* **46** (3), pp. 627–639.

Meystre, S.M., F.J. Friedlin, B.R. South, S. Shen, and M.H. Samore (2010), Automatic deidentification of textual documents in the electronic health record: a review of recent research, *BMC Medical Research Methodology*.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing. http://www.R-project.org.

Rietveld, T. and R. van Hout (2005), *Statistics for language research, analysis of variance*, Mouton de Gruyter, New York.

Shockley, Linda (2003), *Sound Patterns of Spoken English*, Blackwell Publishing.

Warren, P., S. Speer, and A. Schafer (2003), *Wanna*-contraction and prosodic disambiguation in US and NZ, *English.Wellington Working Papers in Linguistics* **15**, pp. 31–49.

Ye, Lu (2005), The stylistic features of webchat English, *Cross-cultural Communication* **1** (2), pp. 73–76.