Published in Revue Tranel (Travaux neuchâtelois de linguistique) 63, 49-69, 2015
which should be used for any reference to this work

Computer-Mediated Communication: What a Quantitative Linguistic Approach Should Do

John C. PAOLILLO
School of Informatics and Computing, Indiana University

Cet article met en perspective une approche quantitative de données linguistiques issues de la Communication Médiée par Ordinateur (CMO) et en propose quelques principes d'application. Les méthodes de cinq approches quantitatives différentes sont passées en revue et discutées tant du point de vue linguistique que statistique. Les observations faites suite à l'examen de ces méthodes permettent de dégager un ensemble de principes à respecter dans l'approche quantitative de données linguistiques. L'article clôt avec quelques conclusions d'ordre général à propos de l'application de cette approche à l'étude de la CMO.

1. Introduction

Linguistics has played an important role among disciplines seeking to understand computerized communication and its consequences. Some of the earliest studies of computer-mediated communication (CMC) draw on linguistics or employ linguistic insights. The scale of the adoption of CMC and the availability of data encourage quantitative analysis. Unfortunately, the number of studies on CMC that are both linguistically informed and quantitative is relatively small. A consequence of this is that large-scale quantitative studies are often done in the absence of linguistic insight, leading to many spurious or incorrect conclusions. Hence, quantitative studies of CMC need to become better informed linguistically, to improve upon our understanding of the phenomenon.

My goal here is to sketch a role for a quantitative linguistic approach to CMC and to indicate what it could look like. In the first part of this discussion I suggest a motivation for a quantitative linguistics of CMC. I then briefly summarize some existing quantitative linguistic perspectives, sketching the linguistic and statistical reasons why these approaches fall short of what is needed. I then offer a set of principles for quantitative linguistic analysis, and close with general conclusions.

2. Why should there be a quantitative linguistics of CMC?

In most parts of the world, information technologies are now tightly integrated into all our patterns of communication. For many people, daily rituals include checking email, logging into social media accounts, texting and sending "selfies" on smart phones, searching the Internet for entertainment, reading online news, and video-chatting with co-workers, family or friends. Many for-

merly print-mediated functions, from commerce to government, healthcare, education, entertainment and recreation are now performed involving some combination of telephones, wireless communication and computers. Businesses routinely create organizational units whose members interact primarily or solely via technology. Outsourcing technologically connects developed and developing regions in a tightly coupled economic network; much outsourced work, such as that in call centers, involves communication. In Africa, cell phones connect remotely located farmers, fishers, producers and craftspeople to information about the markets that they depend on, while in Northern Canada medical care for remote Inuit villages is delivered via video teleconference. Political movements are chronicled in real time on Twitter and Facebook, even if their shape and direction is not directly influenced by those technologies. Many more uses exist for computerized, Internet and telephonic communication than had been originally imagined. CMC is therefore an object of deep concern to everyone inside and outside of academia.

CMC is both ubiquitous and voluminous, making it deeply relevant to linguistics. For many young people, it is even an important vehicle for childhood socialization, with instant messaging, SMS, Twitter and chat having sometimes supplanted the bonding role of face-to-face communication with peers and adults. Longstanding issues within linguistic theory, regarding the relationship of communication modality to language, its use, structure and change, are rendered potentially observable in CMC as never before. Time-scales in language use, of both great depth and fine local detail, are now becoming available for research. The digital nature of CMC also facilitates recording both the content and context of human social interaction; the scales on which this can be accomplished have never before been seen. For these reasons, linguistics can hardly ignore CMC and its theoretical importance.

Moreover, because of the scale of the issues, and the precision required for some questions, quantitative approaches will need to play a role. This implies the application of statistical models,[1] which offer the only rigorous procedure for deciding if one's observations result from the operation of chance, as opposed to systematic and interesting causes. Hence, we must look for insights into how linguistic hypotheses may be expressed quantitatively, and how the models suggested might structure our inquiry into the properties of CMC.

[1] We ignore purely descriptive quantitative analyses for the present purposes.

3. What quantitative linguistics of CMC presently does

There are a small number of well-established approaches to quantitative linguistics, and a number of approaches to quantitative analysis of language from outside of linguistics that are relevant; five are discussed below.[2]

Labovian variationism is the study of language initiated in the 1960s by William Labov and his students and colleagues, which systematically explores language variation through examining fundamental units called *linguistic variables*. Each linguistic variable represents a non-absolute choice among alternatives in some context. Linguistic variables are typically conditioned by a number of different factors, both social (dialect, ethnic group, gender, social class, register, etc.) and linguistic (preceding and following environments, etc.). For analysis, variationists typically use logistic regression models to measure the effects of different factors on the observed choices among alternatives, from which interpretations are made about differences between dialects, social interpretations of the different realizations of a variable, and general processes of language change.

In speech communication, variationist analyses typically focus on phonological and lexical variables, mainly because of the genesis of variationism in urban dialectology, but they can be applied to any kind of structural element where a choice between alternative realizations is involved. The variationist approach has been applied to CMC in a variety of places, such as in the analysis of chat-specific variants (Paolillo 2001), dialect features in chat (Siebenhaar 2006, 2008), genre and gender in weblogs (Herring & Paolillo 2006), among others.

For example, if one is interested in the realization in chat of standard orthographic [s] as non-standard orthographic [z], then the proportion of non-standard [z] (with respect to [s]) will be the *response variable* or *dependent variable*, which we will notate y. Our working hypothesis is that the distribution of the response variable y is influenced by several *predictor variables* (also called *independent variables*) x_1, x_2, etc.; these represent observations about specific contexts of use of [s/z], for example, the linguistic context (word-final or not), the pragmatic function (joking or serious), social identities of the interlocutors (younger users or not), etc. Predictors can be of any data type (categorical or continuous), although their specific handling may depend on data type to some extent. The model statement for a typical variationist model may be given as in (1) (Agresti 1996).

(1) $$f(y) = a + b_1x_1 + b_2x_2 + \ldots + e$$

[2] The selection of approaches is driven by mathematical coherence, as should be evident below. Given space constraints, I cannot address some other important approaches, such as natural language processing (NLP) and literary stylometry.

The function $f(\cdot)$ is a *link function* transforming the observed y values into a scale that can be used in analysis. Typically this is done for value scales like counts or proportions, whose values are limited by zero and/or one; for the logistic regression model used in variationist analysis, the *logit* function is used as $f(\cdot)$.[3] The symbols a and b are model parameters, i.e. values that are to be estimated from the data. The selection of these parameters represents different hypotheses about the distribution of the data, and parameter values and significance tests guide our interpretations of that distribution. Finally, the term e represents the *error* specific to a given observation, but characterized by a probability distribution. This can also be understood as the contribution of chance to the observation.

The model statement in (1) says that the propensity to use a particular variable in a context can be expressed as a simple sum of terms, each of which is the product of a parameter and a corresponding predictor variable (they are multiplied together), except for a, which is the same for all observations, and e, which is distinct for each observation. The propensity can be transformed using the inverse link function $f^{-1}(\cdot)$ to express it in units like proportions.

For interpretation, the interesting parts of the model are a, which can be taken to represent an average propensity of the variable, and the b values, which express the effect of various contextual variables x on the variable of interest y. For example, if one is interested in the realization in chat of standard orthographic [s] as non-standard orthographic [z], then the proportion of non-standard [z] (with respect to [s]) will be the variable y. Predictor variables x_1, x_2, etc. would be observations about specific contexts of use of [s/z], for example, the linguistic context (word-final or not), the pragmatic function (joking or serious), social identities of the interlocutors (younger users or not), etc. The b values indicate how much each contextual factor influences the expression of [s/z], while the a value indicates an overall propensity to use [z] instead of [s].

Commitment to this type of model is faced with several technical and methodological difficulties. The first of these is that linguistic variables must be investigated individually, when the systemic nature of language varieties indicates that different linguistic variables should correlate: [s/z] use is likely to share many contextual characteristics with [are/r] and [you/u] for example (cf. Paolillo 2001). Unfortunately there is no natural way to address this in (1), and multiple such models must be proposed and investigated separately for each variable of interest. This raises a serious statistical problem, as the variation in each variable shared with other variables is unaccounted for and the significance tests for the a and b values are distorted by this.

[3] The logit is the natural logarithm of the odds, i.e. $\operatorname{logit} p = \ln p/[1-p]$.

A second technical problem that worries many analysts is that the variable is typically assumed to be a binary choice among variants of an item. Many relevant phenomena do not have this character, such as the variation among emoticons, or the use of a specific semantic marker (e.g. invariant durative aspectual *be* in African American Vernacular English); treating these in the variationist framework using logistic regression is possible (Rousseau and Sankoff 1978). Technical solutions involve the explication of systems of choice, or adoption of a multinomial logistic regression model for analysis. The latter choice can lead to considerably more effort in interpretation.

A third problem with regression-type models as in (1) is that the practices around their use favor simpler models, sometimes unrealistically simple ones, in a range of ways. First, models with fewer terms are simpler; hence if one doesn't know (or simply doesn't suspect) that some contextual factor influences the rate of variation, one might leave out any term with that factor. Any variation attributable to this factor is therefore available to be associated with other factors, leading the significance tests of those factors to be biased. Recently, a methodological and statistical debate has developed around the treatment of individuals in variationist models (cf. Johnson 2008; Tagliamonte and Baayen 2012; Paolillo 2013) whose central issue how to include a term representing the effect of the individual in a model of the form in (1). Similarly, complex conditional relationships among contextual variables need to be expressed as interaction effects; not only are these difficult to state, but there are many ways to state equivalent effects and models that nonetheless suggest subtly different interpretations. In any case, when interactions are left out, similar problems arise; some of the methodological and statistical issues around this have been explained in Sigley (2003).

Biberian multidimensionalism is an approach established by Douglas Biber in the late 1980s in work with Edward Finegan and other colleagues, focused on revealing dimensions of systematic variation across various contexts of language use, characterized as *registers*, although others prefer the term *genre*.[4] In this approach, language features from a predetermined set are counted in the texts of a corpus, which are grouped into pre-determined categories of communication. The feature-by-category counts are transformed mathematically and subsequently analyzed using Factor Analysis, possibly using a non-orthogonal rotation. The resulting dimensions are interpreted as dimensions along which different message types (registers) vary, according to mode (written/oral), purpose (informational, argumentative, etc.), narrativity, historical period, etc. The multidimensional approach has sometimes been applied to

[4] In this view, genres represent categories of communication (Hymes 1974) like business emails, status updates, picture captions, etc., whereas registers represent language varieties that are specialized in social function. Both types of variation are called "register" variation in the multidimensional approach, in spite of the fact that they can be distinguished both methodologically and theoretically.

CMC (Yates 1996; Emigh & Herring 2005), mainly with the aim of characterizing CMC with respect to speech and written communication.

The model employed in the multidimensional approach may be schematized as in (2). Unlike the model in (1), there are multiple equations in this model, one for each dimension in the result, where s_n represents the *factor score* for dimension n. Also unlike (1), the linguistic features y are considered together; the various *factor coefficients* $b_{n,m}$ relate each of the linguistic variables to the different dimensions of variation. The function $z_n(y_n)$ represents the treatment of the counts, proportions or other measure of the linguistic feature before it is entered into the factor analysis model. Typically this is a *z*-score normalization, although logarithmic or other scale transformations may be applied first. Note that each feature y has its own function $z(\cdot)$. Not represented in (2) is an assumed error term that is specific to each of the linguistic features *y* observed.

(2)
$$s_1 = b_{1,1}z_1(y_1) + b_{1,2}z_2(y_2) + b_{1,3}z_3(y_3) \ldots$$
$$s_2 = b_{2,1}z_1(y_1) + b_{2,2}z_2(y_2) + b_{2,3}z_3(y_3) \ldots$$
$$s_3 = b_{2,1}z_1(y_1) + b_{3,2}z_2(y_2) + b_{3,3}z_3(y_3) \ldots$$

An aspect of the model's application not visible in (2) is that the linguistic features *y* are typically aggregated over some range of features, as well as over some group of example texts. For example, the part of speech category "prepositions" is a language-specific list of words; these alongside all other linguistic features are counted in a corpus of texts that has been partitioned into sets representing non-overlapping "registers", such as conversation, personal letters, business letters, public speeches, etc. This means that aggregation is taking place at two different levels, that of the linguistic feature and that of the register. The results of the analysis are determined in great part by these aggregations in ways that cannot be fully accounted for, as nothing is left to trace back from the aggregate features and registers to the individual features and their specific contexts of use.

Multidimensional analysis is conducted on a sample of texts that is intended to be representative of the relevant range of variation in a particular language or variety. The factor analysis model is highly data-dependent, meaning that very different factor structures may emerge out of only partly different samples. Consequently it is of utmost importance that the contents of the sample are known, and that anything that may lead to observed variation, systematic or spurious, is understood. Generally, these samples are linguistic corpora, sometimes commercially licensed, sometimes purpose-built. Their status as corpora generally means that they are used, often by the same people, in multiple studies. Such sample/corpus re-use appears to be efficient, but in fact it is a deprecated statistical practice, especially when distinct studies appear to show distinct, independently supportable results. This problem is only fully cor-

rected when subsequent studies subsume all of the considerations of prior studies, or when entirely new data is used.

Apart from data re-use, even the largest corpora have serious limitations. For example, in Bresnan and Ford's (2010) study of the English dative alternation in a corpus of voice telephone calls, only 2'360 instances of the dative construction were recorded, from a three million word corpus. Were one to find that this corpus is 100 or 1000 times too small for its purposes, expanding it could only be done at considerable cost. Obtaining samples large enough to show meaningful distributions of rare variables is far from trivial. Worse yet, linguistic variables tend to have highly skewed distributions: high frequencies of occurrence are concentrated in a small number of texts, and rather large numbers of texts show little to no use. This is, of course, the famous Zipf distribution (Zipf 1935; see Baayen 2001 for a more current view). Its consequence is considerable, however, in that it also induces apparent correlations in word frequencies, and it can be difficult to demonstrate that these are not spurious consequences of the chosen sample texts.

Multidimensional research has other difficulties as well. Its features are highly language specific, and under-theorized with respect to other aspects of language, e.g. dialect and social variation, the syntax, semantics and pragmatics of language, etc. It is likely that at least some aspects of the variation observed are due to such unobserved variables in the analysis, and the model does not provide for how these different levels of linguistic analysis should interrelate. Furthermore, some features, such as part of speech tags, are heavily influenced by other factors, so the model sought in multidimensional analysis cannot properly account for the distribution of its selected features, meaning that the resulting dimensions are confounded by these other factors. Consequently, a multidimensional analysis can only be considered broadly suggestive about the nature of the different dimensions of language variation.

The vector space model is a general term we can use to describe a family of approaches derived from the work of Salton in the early 1970s on information retrieval (IR), which is the basis of most search engine technology today. The central idea is to characterize documents in terms of their vocabulary. This is done by counting the frequencies of terms in documents, and arranging them together in a very large term-document matrix. Modern search engines may use up to tens of thousands of terms and millions of documents. The term frequencies are weighted and normalized, using a term-weighting formula, many of which are versions of *tf-idf*, where term frequencies (*tf*) are log transformed and weighted by the inverse of the document frequency (*idf*). Various other operations may be performed to simplify the term-document matrix, such as clustering, dimensionality reduction or a combination thereof.

The vector space model is often not characterized in statistical terms like (1) and (2).[5] Instead, the term-document matrix itself is often treated as the model, and many questions involve the projection of a new document into this "vector space". This is accomplished by counting its term frequencies, weighting them and otherwise mathematically treating them like the original documents to give a term *vector*, i.e. a list of weighted frequency values for each term. This vector can then be compared with the other documents in the vector space using similarity metrics such as the cosine or the Pearson correlation coefficient. In retrieval, documents are ranked according to their similarity to the new document, typically called a *query* and possibly consisting of only a few words. In document classification, the new document's similarity to a number of known document clusters is measured; it is classified as belonging to whichever cluster is closest.

Implicitly, this use of the space with cosine, correlation or Euclidean measures of similarity implies a model like (2), in which the function $z(\cdot)$ simply represents *tf-idf*, and the number of dimensions s_m as well as the number of variables y_n is very large (thousands or millions). One version of the vector space model, known as Latent Semantic Analysis (LSA), has that very description (Landauer & al. 1998; Dumais 2004). This permits some economy in the number of dimensions m that needs to be retained. If one replaces *tf-idf* with a *z*-score normalization, a vector-space model equivalent to (2) results. This has the advantage of being somewhat more motivated from a statistical perspective, and suggests that principal components or factor analysis may be used in place of typically application-specific computation methods, both for obtaining the number of dimensions m, and for computing and interpreting the desired vector space (Paolillo 2004).

There is some ambiguity, however, as to whether terms or documents represent the variables y in the vector space model. The difference in this choice is referred to as P-mode versus Q-mode analysis in the factor analysis literature (Basilevsky 1994; Gorsuch 1983); some versions of $z(\cdot)$ can give identical results for both modes, but that is not the general case and only one arrangement, with terms as y, has a transparent linguistic interpretation.

Apart from this, the notion of document involves inescapable and often arbitrary aggregation over some amount of text; where CMC is concerned, this can be highly unnatural, e.g. when a set of tweets or status updates are aggregated together in order to populate terms in an otherwise very sparse vector. Also, this aggregation is quite different from the kinds of aggregation permitted by (1). In the vector space model, documents are the only factors observed conditioning term distribution. However, the notion is used very elas-

[5] The vector space model is often introduced as a mathematical model, a tacit admission that it lacks fundamental components of a statistical model, such as a random distribution in reference to which significance tests may be made.

tically, and, depending on the application, a document may be an article, an abstract, a paragraph, a sentence or some other unit entirely. All provide very different information about the term and its context of appearance. One does not necessarily know from the presentation of a vector space model exactly what aggregation has been done, or if documents really represent some other contextual variable of interest. Consequently, documents may conflate multiple levels of description, thereby confounding any explanations of a term's distribution.

The vector space model has other problems that make it unsatisfactory as a linguistic model. In any application, there are numerous unobserved factors that also affect term distribution: genres, authors, audiences, etc. A key issue is syntax: there are numerous syntactic and semantic dependencies that are readily observed through conventional linguistic methods, including phrase-structure grammar, lexical subcategorization, selectional restrictions, lexical priming, lexical cohesion, etc. These effects induce correlations among term occurrences in a document (one view being that whatever is intended by "document" is essentially a syntactic unit). Vector space representations are usually insensitive to any word order dependent relations: predication and negation are among these.

The vector space model is also sample-dependent in the same way as multidimensional analysis, and all the same issues attend this. For the vector space model, the corpus is the document set, and very commonly, these are of all one type: journal article abstracts, encyclopedia entries, email messages, status updates, or whatever the researcher has to hand. One can see that quite different results could be obtained from these different kinds of corpora. It is important to recognize that the original purpose of the vector space model is retrieval. For this purpose one wants a representation that arbitrarily closely matches the original document, while being easy to store and to compare against other documents and queries. The term vector suffices for this purpose. It is only later that this has come to be interpreted as a semantic representation (in the sense of "aboutness"), and this was done without carefully considering its statistical consequences, with respect to model structure and sampling needs.

Sentiment mining is an approach to analyzing text that combines aspects of the vector space model with psychological assessment instruments from clinical psychology. Various tools have been designed by different research groups to try to evaluate moods or psychological states represented in psychological assessment interviews. These tools, such as Linguistic Inquiry and Word Count (LIWC; Tausczik & Pennebaker 2010), Affective Norms for English Words (ANEW; Bradley & Lang 1999) and Profile of Mood States (POMS; Bollen & al. 2011), involve a fixed dictionary whose elements are counted in the interview transcript. The counts of the dictionary words are then weighted and summed according to a formula established in prior research of a corpus

of interviews. The results of the formula indicate the "sentiment", "affective content" or "mood state" of the evaluated interview.

The model corresponding to sentiment mining is essentially a vector space model as in (2), but with the linguistic variables trimmed to a small subset of terms that are expected to reveal sentiment information, and the factor dimensions optimized to express some theoretical model of sentiment. It therefore has many of the problems of vector space models, such as insensitivity to syntax, negation, any complex conditions on semantic composition, pragmatic indirectness, etc. The justification for sentiment mining's approach resides in an implicit assumption that dimensions of sentiment are a subset of the various ("aboutness") dimensions of meaning, and that they are independent of and orthogonal to any other dimensions of meaning. Reduction of the space to a subset of its dimensions is treated as a mathematical operation of projecting the vector space into a smaller-dimensional subspace, in the information retrieval literature (Korfhage 1997).

Such assumptions are unsafe, however. For example, if one were to sample Twitter at the time of the downing of the Malaysian airliner MH17 over Ukraine, one would find a range of sentiment-laden terms associated with tweets around that event, especially representing negative sentiments (anger, frustration, confusion, etc.). This would be quite different from the sentiments expressed on another topic, such as the games in the FIFA world cup tournament, the Tour de France, etc. Moreover, in each of these cases people with different orientations will express their sentiments about different things, e.g. both Ukrainian and Russian sympathizers might express anger around MH17, but with opposite targets in the different communities. Sentiment mining cannot address this without substantial changes to the model.

The problem with sentiment mining is two-fold. On the one hand, there is an implicit larger vector space, some of whose terms are potentially correlated with the sentiment terms in unknown ways. Since the larger vector space is unavailable for interpretation, the unobserved terms, and their correlations with the observed terms are also not available, but they nonetheless confound the intended interpretations: variation attributed to affect may well come from another source, such as genre, topic, individual style, etc. On the other hand, many aspects of sentiment are best expressed conditionally: someone is angry *at* a particular event (but happy about something else), etc. Sentiment mining also extrapolates the affective values of terms from a clinical context to an unrestricted general context, in which pragmatic and interpersonal context plays a significant role in fixing affective interpretation. When the context is CMC in the form of blog posts or Twitter feeds, the affective meanings of the sentiment dictionary terms are unlikely to be fixed in the assumed way. Such

complex conditions are not easily expressed in a vector space, which tends to be strictly linear in its composition.[6]

Network analysis is not really a linguistic analysis, but it is a form of quantitative analysis often performed on CMC data, and therefore involves language. Basically, network analysis reduces the meaning of a message to its sender and recipient. Typically, no information about the content is retained, although time/date and other contextual variables may be recorded. Sender and recipient are generally assumed to be the same type of actor, and analyses are conducted to reveal the structure of the "*x* sends *y* a message" relationship among the all the actors. Typically, network visualizations are produced for interpretation alongside various sorts of statistical models may be used: *blockmodels* perform cluster analysis on the network ties to reveal social segmentation (e.g. class, ethnicity, gender, age cohorts, etc.) and flows between different groups (Doreian & al. 2005); *exponential random graph* (ERG) models treat network ties as inter-dependent and seek to reveal relational properties such as reciprocity and transitivity (Handcock & al. 2003); *dynamic process models* treat network ties as generated by a stochastic process and seek to discover process-generating rules which lead to the observed networks (Pastor-Satoras & Vespignani 2007).

Network analysis became popular in early research on CMC because it promised a way to address issues of social interest by directly operating on observational artifacts of CMC: chat log files, email and discussion group archives, etc. could be readily captured and analyzed by largely automatic means to obtain understandings of both social interaction and electronic communication in quantities that were previously prohibitive. Much work came from the perspective of Social Network Analysis (Paolillo 2001; Wellman & al. 1996), which already had a history of developing quantitative, computer-assisted methods for analysis of social interaction of different sorts (Freeman 2004).

At its simplest, the network model is like a vector-space model, where the terms are the various available recipients of messages and documents are the recipient lists of each sender. The analysis performed on this arrangement is often a clustering of senders and recipients, most often with both modes treated at the same time. The general term for this approach is blockmodeling, and it results in a reduced representation of the network that is more readily interpreted than the original network (Doerian & al. 2005). A key observation that is made from this concerns the *centrality* of different senders or recipients, with different definitions of centrality corresponding to different kinds of power or status in the network.

More sophisticated network approaches adopt the same initial arrangement, but develop statistical models where the terms represent elemental network

[6] If a vector space is to be used for sentiment with any accuracy, then the complex conditions associated with the sentiment's target have to be built into the vectors (variables) in some way.

configurations: participants' propensity to link to others, the reciprocity and transitivity of links, etc. This is represented by the ERG approach (Handcock & 2003), and the dynamic process modeling approach (Pastor-Satoras & Vespignani 2007): both result in a summary of global properties that characterize the network at the level of the individual. One example of a network analysis of CMC can be found in Ronen & al. (2014) who examine the status of languages in Twitter and Wikipedia by employing network maps and centrality measurements to arrive at their interpretations.

Network analysis, however, represents an extreme reduction of the message to a single observable: the fact of a message being sent from one individual to another. The message content is seldom accessed, and when it is, it is typically reduced to a small number of categories, to make it amenable to the available statistical models. For example, in the Ronen & al. (2014) study, neither the Twitter nor Wikipedia analyses employed the content of the tweets or articles that "connected" languages. This results in an extreme loss of information about the subject of interest (connections among languages), which constitutes the network in the first place. For example, Ronen & al. (idem) find strong linkages in Twitter between English and three other languages: Malay,[7] Spanish, and Portuguese; a fact which is impossible to interpret without knowing what is actually shared among the relevant groups. And although structures may be observed in the resulting network, we lose the opportunity to examine what *in the messages* might have caused this structure; in other words, the observation of structure in a network analysis of CMC is confounded by the unobserved contents of the communication.[8] Furthermore, all communication networks are observed in some time-window; the choice of that time window has an enormous impact on the nature of the structures observed. For example, with regard to the Ronen & al. (2014) study, Twitter data was only collected from Dec. 6, 2011 to Feb. 13, 2012. Connections made outside this narrow time window are completely unobserved, however important they might be to the interpretation. Hence, sharply reductive procedures like network analysis cannot guarantee a readily interpretable result without additional explanations or methodological constraints.

4. Principles for quantitative analysis in CMC research

It should be evident from the summaries given so far that each form of quantitative analysis of CMC has its own limitations. Each is intended to answer a certain type of question, and makes a set of assumptions amenable to that

---

7 The "Malay" language classification includes Indonesian in Ronen & al. (2014).

8 A related problem is that individuals are treated as equivalent, but for their connections to others. Hence, network models cannot help us see what within the individuals (personal histories, cognitive propensities, etc.) might be responsible for their communications, and thus their connections with others.

goal. Each ignores other types of information relevant to questions that the other approaches may address. None of them addresses all of the relevant questions, nor does any provide a framework for selecting among competing approaches. My goal in this section is to recommend a general set of principles to address these issues and guide quantitative CMC research design.

***Avoid reductionism*.** Reductionism may arise in any of the approaches, although its specific manifestations vary in each. Variationist methodology reduces by focusing on individual linguistic variables, when multiple variables potentially share their variation. The multidimensional approach, although it considers multiple variables simultaneously, nonetheless ignores potential differences among many individual variables, such as specific verbs within the various verb classes. The vector space model ignores relevant syntactic and social conditioning factors, thereby impoverishing its "semantic" representation. Sentiment mining focuses on a specific subset of lexical variables to the exclusion of all others, and network analysis reduces all communication to the identity of its endpoints.

There are at least two manifestations of reductionism in research design: one arises when relevant contextual variables are ignored, leading to incorrect inferences about the phenomenon observed. The other arises when disparate elements are treated together as representing the same variant of a variable. Both of these reductions result in an improper aggregation: elements are counted together when they at least potentially should not be.

The aggregations used in a study must be defensible, on a theoretical level. If they are not or cannot be defended, they imperil the interpretation of the analysis. This sort of issue has gathered a considerable amount of healthy, methodological argument in the variationist literature, e.g. around the treatment of different forms of *be* in English (Rickford & al. 1991) or realizations of /s/ in Spanish (Sankoff & Rousseau 1989). The problem, however, is a general one, and its consequences in the other approaches are less extensively explored. The categories used in a piece of quantitative research, and the aggregations they result in are a critical aspect of the research design, and nothing can be safely interpreted without them.

Hence, careful attention must be paid to how the categories defined aggregate different phenomena together, and the analytical choices made should be carefully defended, based on theoretical and methodological considerations. Methodologically, if models can be constructed that allow the alternative aggregations to be compared, then one can address the aggregation issue as part of one's research questions. This, for example, is the point of work by Sankoff & Rousseau (1989) on rule ordering in Spanish /s/ deletion (see discussion in Paolillo 2002: 93).

Account for skewed distributions. Many of the types of data handled in linguistic analyses have highly skewed distributions. The skewness of linguistic

distributions has a long history of study, going back at least to Zipf (1935). Unfortunately, few of the issues that skewness represents in linguistic distributions have been resolved in any general way. Hence, the consequences of skewness deserve direct consideration in any research design.

There are two ways skewness is manifest in linguistic data: size extremes and sparseness. These raise different issues and they require different solutions. The size extremes of a distribution are closely related to the scale on which it is measured. When categories are observed, the categories are typically counted, and counts are limited at zero, but unbounded in the positive range: no category can occur a negative number of times, but high values can occur, although less frequently. Such distributions are usually transformed logarithmically when analyzed: on the logarithmic scale, both positive and negative values are permitted, with the negative values corresponding to fractional counts. Because values on the logarithmic scale are compressed with respect to the count scale, aggregation can have a distorting effect. Hence, care in the handling of aggregates is justified statistically and mathematically, as well as theoretically.

The sparseness of linguistic distributions refers to the preponderance of zero values. This is easiest to see in the case of vector spaces: overwhelmingly, most entries of a term-document matrix are zero, meaning that a specific word was not observed in a specific document. This is of great concern in social media data, where posts are typically short, e.g. a single sentence or less. Clearly, most of the words of a language will never have a chance to be observed in any given sentence. When this occurs, the counts have little meaning of their own, and only the occurrence (or lack of it) for a given word is important. Sparseness is worsened by logarithmic transformation: zero, logarithmically transformed, is either undefined or negative infinity, and neither value can be used statistically. In information retrieval, the "log+1" transformation is commonly used: one is added to all the counts before the logarithm is taken. This approach results in another mathematical distortion. If a count is in the hundreds (a common word in a long document), adding one changes its logarithm very little, but when a count is small or zero, it changes it much, much more. There is furthermore no principled reason why the value 1 is chosen; one could just as well add 0.5, obtaining a different weighting of zero cells, and hence a different analysis.[9] There is no simple fix for this problem that preserves the count values; at best, they can be truncated to zero or one (absent or present) and analysis can be conducted on those values. Zero-inflated models generally take this approach while retaining a separate model of the counts, conditional on the non-zero cells.

[9] Beyond this, the weighting of zero cells implied by adding one is dependent on the number of rows and columns in the term-document matrix, and hence on the specific application. For these and other reasons, log+1 is not mathematically well-behaved in general.

Identify appropriate statistical models. All statistical models make assumptions about the nature of the data, its distribution, the way that independence is reflected, and what sorts of things are expected to be independent. Quantitative research needs to carefully consider these model assumptions and match them to the research design. If the assumptions do not match the nature of the data, analysis proceeds using an incorrect model, and the results cannot be safely interpreted. Some of the ways that the data distribution bears on model assumptions are discussed above: the value scale and its limits bear on the mode of combination (the logarithmic transformation); the data's sparseness bears on what kinds of values can realistically be used.

Independence of effects is normally reflected by arithmetically adding them together. This is true in each of the models we have considered: the independent terms in (1) and (2) are added together in order to predict some value of interest. Sometimes, however, two variables may be related *conditionally* to the value of interest, in which case they are said to *interact*, or require an *interaction effect*. For example, two users of social media may be inclined to converse in one language, perhaps French, while one of them frequently uses a different language (possibly German) with other interlocutors. The French/German language preference is conditional on the identities of both interlocutors, rather than solely on the propensity of one or the other to use each language. Mathematically, this is handled by creating a term in which the interacting variables are multiplied together, with its own *b* coefficient (the variables must be coded on a scale that allows this; see Paolillo 2002 and Sigley 2003 for discussion).

Other complex terms can also be necessary if the variable has a non-monotonic relationship to some contextual variable. For example, Siebenhaar (2006) observes a common sociolinguistic pattern in real-time chat in which dialect choice is conditioned by age. As often happens, this choice appears to be different in the middle age range (more standard dialect) from the older and younger age ranges (more regional dialect). Age is therefore not monotonically related to dialect choice, and a quadratic term (bx_{age}^2) is one way to express this.[10] Higher-order polynomials may occasionally be justified as well, and periodic effects are common in CMC, which often exhibits daily, weekly, monthly and annual usage fluctuations. These considerations bear directly on the nature and complexity of the terms that should appear in an appropriate statistical model, and hence the model's structure.

Another issue bearing on model structure is the nature of the result required. Multidimensional and vector-space models exist, for example, because it is

[10] A more common alternative is to break the age variable into age groups or cohorts, especially when there is not a lot of information about different ages, e.g. the age cohorts in Siebenhaar (2006) could be used directly. Results from the two approaches can sometimes be interpreted similarly, but they represent different theoretical statements about age, and they make different assumptions about the availability of data.

expected that the answers to their questions require more than one dimension: genres are unlikely to be successfully characterized by a single dimension of variation, and semantics, even in the sense of "aboutness", needs to be able to distinguish many different meanings. If the sequence of linguistic elements is at issue, as it typically is anytime, syntax is involved; the model must be structured to account for this as well. Similarly, hierarchical structure, another aspect of syntax, may need to be part of the model as well. When varying combinations of these concerns are involved, a statistical model can get quite complex.

There are three general ways that models can be structured to meet these various conditions. The simplest is if complex conditional terms can be introduced to account for the required structure. Syntactic sequences sometimes have this characteristic. If hierarchical structure is involved, however, other complications are likely to be required. Cascaded models, in which distinct variables are studied in separate models and one model is conditioned on the outcome of the other model, are sometimes recommended for such situations. The Rousseau & Sankoff (1989) models for rule ordering fall under this approach; other arrangements would be required to address hierarchy in phrase structure.

The third approach is to complexify the model by increasing its dimensionality, as in the multidimensional and vector space models. This approach has limitations, and it is difficult to combine complex conditionals and hierarchical arrangements with vector spaces, which characteristically offer a relatively uniform field of values. Such a combined model, to my knowledge, has never been attempted, and the complexity of the model is one reason for this.

In designing CMC research, it is critical to think through the implications of one's questions in terms of the kinds of relationships that are involved, as this bears on the selection of the proper statistical model to employ. Moreover, one must be responsible not only to one's desired interpretations, but to any other factors that are not of specific interest, but which nonetheless are relevant to the interpretations that could be drawn. For example, Herring & Paolillo (2006) demonstrated that presumed gender effects in weblogs (Koppel & 2002) could be attributed exclusively to genre: failure to include genre effects in the original model results in an incorrect interpretation. This same lesson applies generally to all of the different kinds of effects one must consider. Hence it is crucial to identify, for the sake of properly specifying a statistical model, what effects one is responsible for in an account of CMC variation. The arguments for this come from the research design and its relation to the statistical model, and never from within the statistical model itself, in spite of many assertions to the contrary one can find in the literature.

Answer all research questions in one model. The previous answer leaves us with a big question: how do we design research and choose models to

make the interpretations we would like? If we have multiple questions, can we have multiple models to answer these questions? While this approach is often followed, the short answer is no, it is never safe to interpret multiple models in answering different questions. It is only ever safe to answer all the relevant questions within a single model. The reason for this is technical: when factors are unaccounted for in a model, they may still have effects. They may cause effects included in the model to appear significant, if those are correlated with the excluded effects, or, on the contrary, they may cause them to appear non significant, if they are uncorrelated or anti-correlated. Hence, if effects are excluded from the model that should not have been, alternative interpretations invoking them are fair game.

There are two approaches to addressing this general problem. One requires careful consideration of the various different possible effects and excluding them on the basis of some well-founded theoretical grounds. This is difficult to do, and different members of the field often have different opinions about what is relevant, for different reasons. It is not possible to settle all of the questions one would like this way, and those that are a matter of substantive empirical dispute simply must be included in one's design. Statistically, the criterion of what to include is known as *ignorability*; though this term needs to be understood technically within statistics, its actual meaning is vague and its application to specific factors is subject to theoretical argument in any given study.

A second approach is to employ *controls*, i.e. observational procedures that explicitly take into account some factor, possibly by fixing it to one or more values. If one uses multiple values of a factor as control, of course, then one needs a term in the model for that factor. If one limits oneself to fixing a factor at a specific value, then one's interpretations are limited. They are effectively *conditioned* on that factor value, and the main consequence of this is that we know nothing about what would happen if we allowed that factor to vary. Care must be taken to ensure that the questions asked can tolerate this absence of information.

The observations above have consequences for data re-use, especially the use of corpora, chief among which is that one cannot restudy existing corpora without wanting to either replace or augment the findings of earlier studies, no matter what they are. Since data for a corpus are often collected with specific questions in mind, the information that one needs to answer one's questions is often not available. Consequently, data re-use probably should not be encouraged to the extent it currently is (e.g. consider the number of studies of the Enron corpus, given its availability).

At times, it may be very difficult to answer all one's questions in a single model. For example, suppose one has a syntactic variable that is syntactically and socially conditioned in a way that indexes the identities of both participants. Such a variable could be codeswitching in a mixed-competency bilingual set-

ting. The variable depends on the syntactic environment to license it, so sequence and hierarchy may need to be part of the model. Speaker and addressee need to be known as well, and distinct combinations are relevant to how switching occurs. Topic and semantic domain are well-known correlates of codeswitching as well. The statistical model for this potentially requires parameters for hierarchical and serial syntactic factors, the speaker-addressee network and possibly some unknown number of semantic dimensions. Handling such a model is a worthy endeavor, but very difficult because of the many unsolved problems it invokes and its untested nature. More fruitful work is likely to be done by restricting the questions in some way, e.g. conditioning on the specific pairs of interlocutors conversing within specific semantic domains, and exploring the syntactic properties of the switch, or some other restricted combination of the available variables. This approach sacrifices generalizability, but with the aim of permitting inferences that are sound.

Ensure data sufficiency. A final and persistent hazard of all language-related variation research, and therefore of CMC research, is that there are often far more questions that can be answered than the data permits. This points to lack of successful argument for ignoring irrelevant factors, or the failure to implement meaningful controls, but this circumstance is so ubiquitous and its consequences are so important that it deserves special mention. Relatively simple quantitative analyses require small amounts of data to establish "significance"; larger models require more data. The problem is that the data requirements are multiplicative of the model complexity: adding a factor with three levels requires three times more data to estimate than the model without the added factor. This requirement, though true of all types of models, is widely ignored, much to the hazard of the research.

A relatively prominent example may serve to make the general point. Bresnan and Ford (2010) use a corpus of transcribed English telephone conversations to estimate a corpus model for the dative alternation in English. This model considers nine factors with two levels each (structural paralellism, syntactic complexity, discourse accessibility, definiteness, pronominality, animacy, concreteness, person, number), in examples from 50 different lexemes that are expected to exhibit the dative alternation, in which an indirect object may appear as a bare NP object immediately after the verb, or a PP object with the preposition *to* or *for* in a later position in the sentence. The model therefore implies 512 possible combinations of factors, ignoring the random effects.[11] With only 2'360 data tokens in the telephone corpus dataset, this gives an un-

[11] Considering verb lexeme, there is a total of more than 38'000 cells; the reviewers disagree with my view that the random effects should be considered in the research design, and that the corpus model is under-determined by the data in the extreme. Yet merely including the random effect at its nominal one degree of freedom doubles the effective number of cells to 1024, and halves the average cell count to 2.3 tokens, underscoring the overall problem of data insufficiency.

impressive average of 4.6 tokens per cell, and given the characteristic imbalances associated with naturalistic data, it is highly likely that at least some cells to have no data associated with them.

A closer look at the data reveals that the data for the different verb senses are quite unbalanced. This is what should be expected given the skewed distribution of lexical frequencies discussed above, but its consequence is that any significant factors observed cannot be trusted, because the random effect for verb is unreliably estimated for most of the verbs in the corpus. This situation is normally addressed in variationist research by confining the variable of interest to, e.g. sentences in the corpus in which *give* is the verb of interest (1'666 examples; more than half of the data). This leads to a different kind of study, which is not able to generalize beyond the specific verb *give* (cp. the variationist approach to English *t/d*-deletion, which originated from an attempt to more generally examine consonant cluster reduction).

Data quantity is therefore a paramount concern, which should temper the goals of the research with the cost of data gathering and the complexity of selecting an appropriate statistical model. Unfortunately, the large quantities of data available for quantitative CMC research often lack important contextual information or adequate controls in the research design that would license their interpretation. Worse, such large volumes of data may yet be insufficient if all the relevant research questions were introduced into a single model. This is only addressed by restricting the scope of the research question, carefully constructing the research design, controlling and arguing for the exclusion of factors that need to be ignored.

5. Conclusion

This discussion has emphasized the interrelation between the design of quantitative CMC research and the choice of statistical models that are used in analysis. In addition, the linguistic nature of the phenomena observed has an important status. Much CMC research emphasizes conception of the research goals in model selection; while this is important, it is secondary to the other three considerations. CMC research that uses inadequate statistical models might be redeemed to the extent that it is executed with sound linguistic reasoning and observation. There is little chance that, however, thorough the statistics, CMC research based on shoddy linguistic reasoning can be redeemed. Research design, which variables should be observed, which should be controlled and how, how much data should be collected, etc. follows from the best understanding of the phenomenon studied. For CMC research, in which language plays a crucial role, both the design and the selection of the model need to be deeply linguistically informed.

References

Agresti, A. (1996): An Introduction to Categorical Data Analysis. New York (Wiley).

Androutsopoulos, J. (2011): From variation to heteroglossia in the study of computer-mediated discourse. In: Thurlow, C. & Mroczek, K. R. (eds.): Digital Discourse: Language in the new media. New York (Oxford University Press), 277-298.

Baayen, R. H. (2001): Word Frequency Distributions. Dordrecht (Kluwer Academic Publishers).

Basilevsky, A. T. (2009): Statistical Factor Analysis and Related Methods: Theory and Applications. New York (John Wiley & Sons).

Biber, D. (2006): University language: A Corpus-Based Study of Spoken and Written Registers. Amsterdam (John Benjamins).

Bollen, J., Mao, H. & Pepe, A. (2011): Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: ICWSM 11, Barcelona, Spain, July 2011.

Bradley, M. M. & Lang, P. J. (1999): Affective norms for English words (ANEW): Instruction manual and affective ratings (Technical Report C-1). Gainesville, FL (University of Florida, The Center for Research in Psychophysiology).

Bresnan, J. & Ford, M. (2010): Predicting syntax: Processing dative constructions in American and Australian varieties of English. In: Language, 86 (1), 168-213.

Doreian, P., Batagelj, V. & Ferligoj, A. (2005): Generalized Blockmodeling. New York (Cambridge University Press).

Dumais, S. T. (2004): Latent semantic analysis. In: Annual Review of Information Science And Technology, 38 (1), 188-230.

Emigh, W. & Herring, S. C. (2005): Collaborative authoring on the Web: A genre analysis of online encyclopedias. In: Proceedings of the 38th Hawaii International Conference on System Sciences. Los Alamitos, CA (IEEE Press).

Freeman, L. C. (2004): The Development of Social Network Analysis: A Study in the Sociology of Science. Vancouver (Empirical Press).

Gorsuch, R. L. (1983): Factor Analysis (second edition). New Jersey (Lawrence Erlbaum Associates).

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M. & Morris, M. (2003): statnet: Software tools for the Statistical Modeling of Network Data. Seattle, WA. URL http://statnetproject.org

Herring, S. C. & Paolillo, J. C. (2006): Gender and genre variation in weblogs. In: Journal of Sociolinguistics, 10 (4), 439-459.

Hymes, D. (1974): Foundations in sociolinguistics: An ethnographic approach. Philadelphia (University of Pennsylvania Press).

Johnson, D. E. (2009): Getting off the GoldVarb Standard: Introducing Rbrul for Mixed□Effects Variable Rule Analysis. In: Language and Linguistics Compass, 3 (1), 359-383.

Koppel, M., Argamon, S. & Shimoni, A. R. (2002): Automatically categorizing written texts by author gender. In: Literary and Linguistic Computing, 17(4), 401-412.

Korfhage, R. R. (1997): Information storage and retrieval. New York (Wiley).

Landauer, T. K., Foltz, P. W. & Laham, D. (1998): An introduction to latent semantic analysis. Discourse Processes, 25 (2-3), 259-284.

McNair, D. M., Droppleman, L. F. & Lorr, M. (1992): Edits manual for the profile of mood states: POMS. San Diego, CA (Educational and Industrial Testing Service).

Paolillo, J. C. (2001): Language variation on Internet Relay Chat: A social network approach. In: Journal of Sociolinguistics, 5 (2), 180-213.

— (2002): Analyzing Linguistic Variation: Statistical Models and Methods. Stanford, CA (CSLI Publications).

— (2004): Latent Structure Analysis: Semantic or Syntactic? In: International Conference on Natural Language Processing. Hyderabad, India.

— (2013): Individual effects in variation analysis: Model, software, and research design. In: Language Variation and Change, 25 (01), 89-118.

Pastor-Satorras, R. & Vespignani, A. (2007): Evolution and Structure of the Internet: A Statistical Physics Approach. New York (Cambridge University Press).

Rickford, J. R., Ball, A., Blake, R., Jackson, R. & Martin, N. (1991): Rappin on the copula coffin: Theoretical and methodological issues in the analysis of copula variation in African-American Vernacular English. In: Language Variation and Change, 3 (01), 103-132.

Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S. & Hidalgo, C. A. (2014): Links that speak: The global language network and its association with global fame. In: Proceedings of the National Academy of Sciences, 111 (52), 5616-5622.

Salton, G. Wong, A. & Yang, C. S. (1975): A vector space model for automatic indexing. In: Communications of the ACM, 18 (11), 613-620.

Sankoff, D. (1978): Linguistic Variation: Models and Methods. New York (Academic Press).

Sankoff, D. & Rousseau, P. (1989): Statistical evidence for rule ordering. In: Language Variation and Change, 1 (01), 1-18.

Siebenhaar, B. (2006): Code choice and code switching in Swiss German Internet Relay Chat rooms. In: Journal of Sociolinguistics, 10 (4), 481-506.

— (2008): Quantitative approaches to linguistic variation in IRC: implications for qualitative research. In: Language@Internet, 5.

Sigley, R. (2003): The importance of interaction effects. In: Language Variation and Change, 15 (2), 227-253.

Tagliamonte, S. A. & Baayen, R. H. (2012): Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. In: Language Variation and Change, 24 (02), 135-178.

Tausczik, Y. R. & Pennebaker, J. W. (2010): The psychological meaning of words: LIWC and computerized text analysis methods. In: Journal of Language and Social Psychology, 29 (1), 24-54.

Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M. & Haythornthwaite, C. (1996): Computer networks as social networks: Collaborative work, telework, and virtual community. In: Annual Review of Sociology, 213-238.

Yates, S. J. (1996): Oral and written linguistic aspects of computer conferencing. In: Herring, S. C. (ed.), Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives. Amsterdam (John Benjamins), 29-46.

Zipf, G. K. (1935): The Psycho-Biology of Language. Boston (Houghton-Mifflin).