THEORIA ET HISTORIA SCIENTIARUM, VOL. XI
Ed. Nicolaus Copernicus University 2014

First published online 31.01.2015 http://dx.doi.org/10.12775/ths-2014-003

Maciej Pokornowski

Center for Language Evolution Studies (CLES); Department of English
Nicolaus Copernicus University
pokornow@doktorant.umk.pl

The fourth V, as in evolution: How evolutionary linguistics can contribute to data science

Abstract. The paper explores the importance of closer interaction between data science and evolutionary linguistics, pointing to the potential benefits for both disciplines. In the context of big data, the microblogging social networking service – Twitter – can be treated as a source of empirical input for analyses in the field of language evolution. In an attempt to utilize this kind of disciplinary interplay, I propose a model, which constitutes an adaptation of the Iterated Learning framework, for investigating the glossogenetic evolution of sublanguages.

Keywords: Data science, evolutionary linguistics, natural language processing, Twitter, glossogeny, Iterated Learning framework

1. Introduction

This paper shows the importance of closer interaction between data science and evolutionary linguistics, pointing to the potential benefits for both disciplines. To substantiate the claim concerning these profits, a model is put forward for investigating the glossogenetic evolution of sublanguages emerging on social networking services (SNS), such as Twitter.

Users of social networking services generate unprecedented amounts of text, which constitutes one of the foundational factors of the big data revolution. For linguists this presents both new opportunities and challenges.

On the one hand, these data can be viewed as corpora that are especially valuable if one is to consider their size and actuality. On the other hand, challenges surface, as it is one of the roles of language experts to provide insight that will enable developing tools for efficient execution of natural language processing (NLP) tasks. The optimization of these tasks constitutes one of the principal demands in the emerging field of data science. Despite the relatively satisfactory state of syntactic processing tools, the semantic and pragmatic analyses of texts are still in their infancy (Cambria and White 2014). Similarly, tasks associated with language dynamics await scientific tackling in order to bring algorithms closer to natural language understanding.

2. The three V's of Big Data

Recent developments in digital technology, including the virtually ubiquitous access to the Internet via mobile devices, coupled with the advent of social media platforms, lead to the generation of unparalleled amounts of data. Many intuitive interpretations of the buzzword *big data* overlook the significance of the phenomenon in question. The key aspect is the results that can be achieved through analyses conducted over these enormous sets of information, hence the demand for specialists able to work within the emerging academic discipline, *i.e.* data science. Its inherent interdisciplinary nature renders it rather difficult to be defined unequivocally (Provost and Fawcett 2013: 2–3). In a similar vein, the skills and expertise expected from a "data scientist" usually exceed competencies of any single person; the minimum requirement seems to be the ability to develop codes, operate data analytical tools and, at the same time, have a solid background in the relevant domain (e.g. medicine, marketing, etc.), which defines the type of data for analysis (Davenport and Patil 2012: 73–74). Since at present there are not many academic institutions that offer degrees in data science (Davenport and Patil 2012: 74), the only currently viable solution is forming interdisciplinary groups of experts with appropriate labor distribution. Indeed, big data analyses drive progress in many diverse domains of social activity, including: health care, science, politics, social engineering, economy, and logistics (Conte *et al.* 2014: 326). This demand opens up a niche also for linguists.

In its contemporary interdisciplinary context (Jackendoff 2002, 2007a, 2007b), linguistics plays a pivotal role in the numerous challenges that lie ahead of computational social science. After all, in 2007, a third of all digital data comprised of text (Hilbert 2014). Thus, mainly via the applied end of its disciplinary spectrum, namely natural language processing, the study

of language provides important base-knowledge for conducting analyses on big text data. In general, digitalized language can be considered one important type of structured data, *i.e.* data whose analysis is possible with the currently available tools, such as key word recognition or parsing. However, language data interpretability depends on the particular NLP task and so, for instance, sentiment analysis, relation extraction or textual entailment still constitute computational operations that yield low levels of accuracy, being associated with semantic and pragmatic processing (Cambria and White 2014: 55–56). Looking from the opposite perspective, great amounts of annotated, authentic texts have been the holy grail of linguists (at least, in the usage-based or empirically-driven approaches) for many years now. Thus, processing big data should be understood as having access to basically limitless amounts of information (provided, of course, that the processing yields correct metadata tagging), enabling empirical research that is superior to "rules and logic [that] miss frequency and language dynamics" (Bengfort 2013).

As any relatively new term entering the scientific discourse, big data has been causing some controversy in terms of definition. In a brief review of some of the most important definitions available, Chen *et al.* point out that explanations pertaining exclusively to size are insufficient since they always remain relative to the particular domain of research (consider the sizes of sets in fields such as: genetic sequencing, astronomy, or social science) (2014: 173). Hence the importance of focusing on the remaining two aspects: velocity and variety – following the commonly quoted Doug Laney's "3Vs model" (see Table 1). The authors of the review propose the following synthesis: "In general, big data shall mean the datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time" (Chen *et al.* 2014: 173), due to either their volume and/or variety.

Table 1. *The three aspects of big data recognized in Laney's model*

Aspect	Meaning
Volume	scale and size of data sets, relative to a particular domain and to the computational power required to process them that are available to a particular processing entity
Velocity	pace at which data is produced and the speed required for its efficient processing
Variety	data type diversity in a given stream (text, video, audio, static image, etc.); also differences in data processability(structured, semi-structured, unstructured data)

3. The missing V

In spite of its great potential, data science is still at an early stage of its development as an independent discipline, and there are as many challenges to face as there are opportunities. In a recent review paper, Jagadish *et al.* (2014) identify some of the most significant problems for the developing field: heterogeneity, inconsistency and incompleteness, scale, timeliness, privacy and ownership, human cooperation and comprehension. Some of these issues, for instance heterogeneity, relate directly to problems of language processing if one is to consider features such as diversity of natural languages, genres, spelling conventions, and levels of linguistic analysis. Providing a comprehensive report on the state of the art in the NLP research, Cambria and White (2014) identify the tasks that lie ahead of big-data-driven language processing, that is interpreting texts beyond the purely syntactic level and focusing on the much in-demand semantic and pragmatic information in order to execute: sentiment analysis, emotion recognition, relation extraction, linguistic summarization, knowledge representation, word sense disambiguation, co-reference resolution, question answering and grammatical evolution. The successful management of these NLP tasks will render big language data structured across all levels of analysis, and thus, ultimately make it usable in the context of data science.

To gain a more comprehensive perspective on the types of problems occurring in the field of data science, let us turn to one of the most commonly quoted examples of big data implementations: Google Flu Trends (GFT). GFT is an algorithm developed to conduct spatiotemporal predictions of influenza pandemics based on analyses of Internet users' queries typed into the Google search engine. The accuracy of GTF has been recently questioned by Lazer *et al.* (2014) in a paper that reports significant discrepancies between the algorithm's predictions and factual data obtained from public health care institutions. The authors conclude that the progressing miscalculations in GFT prognoses are a direct result of Google engineers underestimating the dynamics that govern both users' behavior and the technologies/platforms being used:

> Twitter, Facebook, Google and the Internet more generally are constantly changing because of the actions of millions of engineers and consumers. Researchers need a better understanding of how these changes occur over time. Scientists need to replicate findings using these data sources across time and using other data sources to

> ensure that they are observing robust patterns and not evanescent trends. […] More generally, studying the **evolution** of socio-technical systems embedded in our societies is intrinsically important and worthy of study. (Lazer *et al.* 2014: 1205; bold fonts: MP)

Hence the importance of taking into consideration the fourth V, i.e. evolution, when conducting any kind of big data analyses. Language, being one of the systems mentioned above, requires investigation that would embrace its dynamic nature manifesting itself within a relevant time-frame. This is the task, whose accomplishment can be brought closer through the insights stemming from the Evolutionary Linguistics research.

4. Selecting appropriate data sources

One of the key factors affecting any given set is the source of data. Let us consider a question: which platform should one use for data crawling to achieve highest accuracy? After all, search engine queries, SNS posts or microblogging entries differ in terms of form as a direct result of users' interactions with various functionalities provided by a particular platform.

This platform-output dependency can be illustrated with reference to the GFT discussion. In their response to the GFT critique mentioned above, Broniatowski *et al.* (2014) quote their own influenza surveillance studies that were significantly more successful in terms of prediction accuracy while relying on data from Twitter. The authors claim that the main advantage of Twitter data over Google's is the lack of replicability constraint: the latter are proprietary, which restricts any research based on the query data only to scholars employed directly by the Mountain View giant, whereas the former are provided as part of the Twitter's open access policy (Broniatowski *et al.* 2014), which stimulates unmonopolized scientific progress. In fact, the Twitter-based prognoses came satisfactorily close to the actual flu metrics compared against the figures obtained from public health care institutions – an effect stressing the importance of source choice in big data analyses.

5. Twitter characteristics[1]

Twitter is a microblogging social network service that enables its users to post short text entries, called *tweets*, of maximum length not exceeding 140 characters, optionally including also image attachments and/or URLs. When considering the contemporary Social Networking Services with the highest populations, Twitter has a number of features that render it particularly interesting for researchers working in Social Network Analysis (SNA). Firstly, unlike, for instance, Facebook, Twitter does not impose reciprocity in user relations. In other words, user A can follow user B's posts without user B being obliged to follow user A back. Secondly, contrary to the majority of social media, whose evolution is driven solely by their developers, with innovations being introduced in a top-down fashion, Twitter functionalities originate spontaneously as community-based implementations, hardcoded into the platform only once they surface as users' adaptive behavior (Bruns and Burgess 2011: 2). Apart from the original length constraint, virtually all of the now-established Twitter functionalities emerged in a bottom-up manner. The driving force behind such innovations was interaction enhancement. This way, users of Twitter adopted a number of in-text markers that annotate tweets with specific discursive functions. Table 2 explains the canonical usage of these markers; example tweets, carrying a particular tag, are also provided.

The use of the above markers affects Twitter streams leading to the emergence of diverse information diffusion chains. Two main types of communicative systems are recognized on Twitter: hashtag-based topical discussions and community (follower-followee) conversations (Rossi and Magnani 2012: 563; Bruns and Burgess 2011: 6). Each of these communicative networks exhibit different properties.

[1] An in-depth presentation and discussion of all Twitter functionalities lies beyond the scope of this paper. However, it has been already investigated and exposed in numerous sources; thus, I shall focus only on the features that are relevant for the current discussion. For detailed information on Twitter functionalities, I redirect the reader to the Twitter Glossary Section (http://goo.gl/daqhAp) as well as Kwak *et al.* 2010 and Liu *et al.* 2014: perhaps the two most comprehensive studies on Twitter so far.

Table 2. *Markers used in tweets*

Name	Marker	Function	Example tweet
Hashtag	#	marks a tweet as belonging to a discussion on a particular topic	The moment #EuroMaidan started. Late afternoon, November 21st 2013. I got there a bit later, it was dark already...
Reply	@user_name (initial position)	a public message addressed to a particular user	@Fluid_trn I am looking into it now, i haven't found him yet, but if i do find him you will be the first to know :-)
Mention	@user_name (non-initial position)	mentioning a user; not necessarily for communicative reasons	LIVE NOW: @rustyrockets and @ABFalecbaldwin talk about their kiss on @ KeiserReport WATCH HERE: http://rt.com/on-air/
Retweet	RT	quotation; indicates that the content of a tweet is a direct replication of another tweet (adding one's own comment is optional or depends on the available space left)	"RT @mlcalderone: 48 journalists attacked this month in Ukraine demonstrations; over 100 this year: http://bit.ly/19sw6u0"

Community networks, where users are linked via the "follow relations", constitute a prototypical mode of conversation available on Twitter. This means that one user can directly address another user or a number of users. If the addressees reply, a conversation unfolds, exhibiting features expected to be found in most instances of Computer Mediated Communication (CMC). Honeycutt and Herring (2009) corroborate this view: their study confirms that, despite its design function as a microblogging service, Twitter has evolved as a platform capable of hosting conversation. One of the most comprehensive analyses of Twitter so far, conducted by Liu *et al.* (2014) (data sample: 37 billion tweets gathered over a period of 7 years) further proves the conversational capabilities of Twitter. The authors report that the @ marker, used for addressing, is present in about 50% of all contemporary tweets. Interestingly, although Twitter can potentially host conversations with as many as ten participants (Honeycutt and Herring 2009) or more, Macskassy shows that, within his data sample, 92% of all conversational interactions were between two users only and that there is a tendency for responsiveness to decrease as the number of conversation participants grows (2012: 231). Finally, in terms of graph generation, community networks also

exhibit diversity between themselves depending on the particular use of the @ marker: mention *vs*. reply (Cogan *et al.* 2012).

Topical networks, on the other hand, conglomerate around a particular hashtag instead of in between user interaction. Hashtags designate topics, which usually appear in relation to events in the external reality, whether in anticipatory, *ad hoc* or *post hoc* manner (Bruns and Burgess 2011: 7). Crucially, unlike in the case of @-based conversations, the users involved in hashtag discussions do not have to maintain follower-followee relations. Bruns and Burgess note that such topical networks do not form communities:

> The term 'community', in our present context, would imply that hashtag participants share specific interests, are aware of, and are deliberately engaging with one another, which may not always be the case; indeed, at their simplest, hashtags are merely a search-based mechanism for collating all tweets sharing a specific textual attribute, without any implication that individual messages are responding to one another. (Bruns and Burgess 2011:5)

The two types of communicative networks on Twitter also differ in term of size. Depending on a particular hashtag, the topical discussions can generate up to millions of tweets, involving a comparable amount of users (Kwak *et al.* 2010: 597). By contrast, Honeycutt and Herring (2009) showed that, within their sample, the number of tweet exchanges per conversation (user-to-user interactions) ranged 2–30, for participant number ranging 2–10. Of course, conversations between users can also carry hashtags, which mark personal-level debates within large-scale topical discussions (Bruns and Burgess 2011: 4).

Apart from their prototypical use as topical markers, hashtags are also used for emphasis (highlighting keywords), emotive expression (as Internet memes, e.g. #facepalm), or backchannelling (Bruns and Burgess 2011: 3–5). The latter refers to a situation when Twitter serves as an official communicative platform during a particular event. For such occasions, a special hashtag is coined by the event organizers and then propagated among the participants, or anyone interested in the happening, to coordinate and gather all event-related communication. This solution is used mostly for broadcast events, where the viewers are encouraged to take part in a discussion (e.g. #xfactor) or during conferences, as means of information exchange (see Letierce *et al.* 2010 and Weller 2011).

6. Related work – Twitter and Evolution of Language

While the linguistic evolution of Twitter content *per se* has not been investigated so far, related dynamic phenomena occurring on the platform did receive academic attention. These analyses encompass patterns of information diffusion (Kwak *et al.* 2010) or the evolution of the service itself, *i.e.* available functionalities, users' behavior, network structures (Liu *et al.* 2014).

To my knowledge, the paper by Cunha *et al.* (2011) constitutes the only attempt to approach questions of language evolution in relation to Twitter content. Although not explicitly referring to a particular framework or model in evolutionary linguistics, the researchers investigate the adaptive behavior of hashtags. When a new topic appears, Twitter users are free to coin a suitable hash marker to tag their tweet as being a contribution to a discussion on the particular event or theme[2]. This unconstrained production of tags usually leads to a situation where, initially, a number of competing hashtags are in use simultaneously. Later, through competitive mechanisms, the most successful markers prevail and conglomerate most of the topic-related content. The authors report that there is a set of patterns that govern hashtag evolution. One such regularity is the correlation between the number of characters used for a given hashtag and its success – the tendency being that longer hashtags exhibit lower popularity/prevalence. Another insight is that Twitter hashtag evolution follows the "rich-get-richer" pattern: "in some stems, the popularity of the most common items tends to increase faster than the popularity of the less common ones. It generates a further spread of the forms that achieve a certain prestige." (Cunha *et al.* 2011: 61). Although indicating an interesting direction in the evolutionary study of Twitter content, for now, the work is limited to the # markers only; also it does not constitute a strictly linguistic (morpho-syntactic) investigation of the problem, focusing solely on length and orthography.

Also worth mentioning is that Zappavigna (2011b) presents modes of semantic data visualizations for Twitter input, in which she refers to logogenesis, ontogenesis and phylogenesis as part of the adopted terminology. Although the immediate connotations are those with the field of evolutionary linguistics (see Hurford 1990), here, the terminological overlap is only superficial: the author represents the school of Systemic

[2] The only exceptions are situations when there is an official hashtag (see backchannelling, section 5).

Functional Linguistics and uses these concepts as defined by Halliday (see Halliday 1993 in Zappavigna 2011b).

7. The Iterated learning framework

The Iterated Learning framework (IL), largely developed by Simon Kirby, aims at investigating language evolution at the level of glossogeny, *i.e.* the evolutionary changes in the general structure of the communicative code; changes of cultural nature, where the replicators constitute arbitrarily defined linguistic units (Wacewicz 2013: 1). The term *Iterated Learning* refers to a process that is argued to govern, among other phenomena, the cultural transmission of human language: this, in turn, drives the changes occurring in the linguistic code itself (Kirby *et al.* 2014: 108–109). In the process "an individual acquires a behavior by observing a similar behavior in another individual who acquired it in the same way" (Kirby, Cornish, and Smith, 2008: 10681). Through computational and mathematical modelling as well as psychological experiments that simulate the process of cultural transmission *via* iterated learning, Kirby and his colleagues investigate the evolution of simplified language-like codes, or *evolects*[3] (see Jasiński in preparation). By means of generalization over the data obtained, the IL researchers draw conclusions regarding the glossogenetic evolution of human language. In general, the framework constitutes a valuable tool for acquiring data for the study of the adaptive dynamics of language, thus, confronting one of the main challenges in evolutionary linguistics: the scarcity of empirical evidence.

Over a decade of IL research points to the fact that it is the linguistic codes themselves, rather than their users, that adapt towards compositionality and structure. Across all experimental settings (that is, in computer and mathematical modelling, but also with human agents) a similar pattern prevails: the input code, unstructured and comprising of randomized form-meaning mappings, becomes increasingly regular and structured after a specific number of iterations (in other words, transmissions or generations) (Kirby *et al.* 2014).

The proposed explanation for the adaptive mechanisms observable in IL research is that, in this way, the codes react to the narrow bottlenecks they encounter on their evolutionary path (Kirby *et al.* 2014: 108–109).

[3] Experimental artificial languages or mathematical constructs used for modelling processes of structure emergence and evolution. (Jasiński: personal communication)

In a nutshell, bottlenecks constitute various constraints that mediate inter-generation transmissions (Kirby *et al.* 2004; Kirby and Hurford 2002), rendering the post-transmission state of a given code different from its pre-transmission state. Naturally, the process repeats itself with every consequent transmission. In the context of natural language, these constraints correspond to aspects such as the "poverty of stimulus" problem in learning or the human limited working memory capacity in processing.

8. Evolutionary perspective in microblogging discussions

If the findings discussed above are correct, it would be interesting to investigate analogous phenomena in language data of higher resolution. Therefore, I propose a model for studying glossogeny at the level of topic-specific sublanguages emerging on Twitter. Although the model refers to the Iterated Learning framework in a number of theoretical assumptions, it has different focus, methodology and aims.

While the IL framework focuses on investigating glossogenetic phenomena in evolects, the proposed model approaches similar questions, yet, in the context of authentic language data. Twitter topical discussions provide a convenient source of naturally occurring data, wherein topic-specific sublanguages[4] can be extracted via a particular hashtag. As noted above, such topical discussions constitute pools of time-bound utterances relating to a particular theme, yet, by default, not being a result of conversations within communities. Crucially, rather than being a variety of conversational analysis in the CMC context, this approach allows for the investigation of sublanguages, evolving over time as micro-systems, in abstraction, to a certain degree, from the speakers involved. Also, the choice of topical discussions for analysis, significantly increases the sizes of data sets. However, the shift of focus from evolects to natural language is a profound one: we move from dealing with emerging artificial constructs to sublanguages originating within already fully-developed, highly complex linguistic systems.

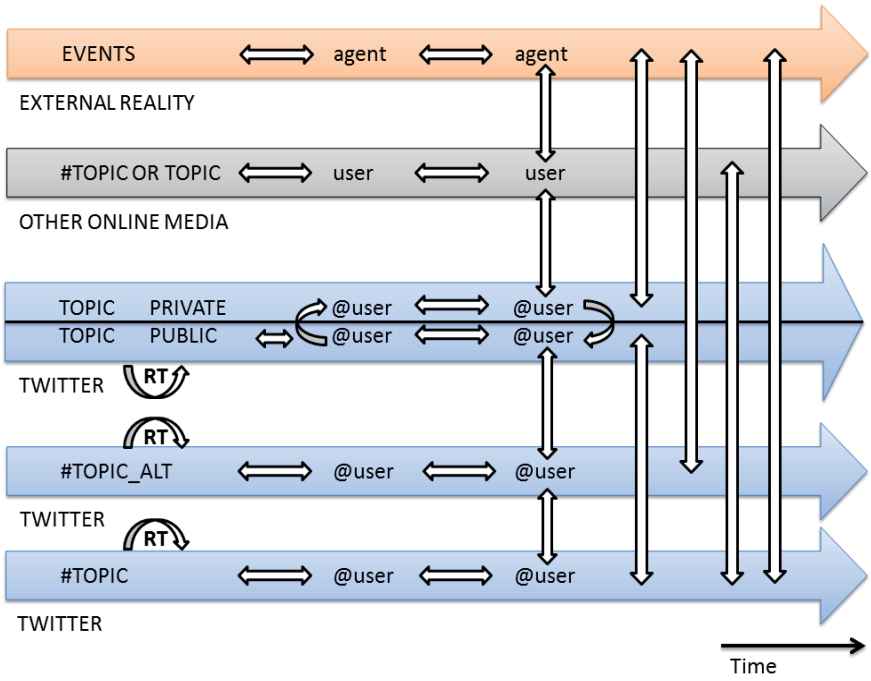
As in the IL framework, the proposed model also assumes a number of bottlenecks on the process of code transmission. The transmission

[4] Though initially I opted to use the term "sociolect" as semantically closest to what is meant here, its definition implies the existence of a particular community using such a variety of language, which could be misleading in the context of topical discussions on Twitter, see section 5.

bottleneck(s) should be understood as a set of constraints that limit the number of possible features which would otherwise freely prevail throughout the evolution of a given code. The expected bottlenecks can be classified as relating to one of the three relevant aspects of Twitter discourse: content, platform and agents. Figure 1 presents a portion of all possible pressures and biases that affect an emerging sublanguage on Twitter.

One bottleneck relating to content is the topic-hashtag association: any given topic is normally discussed under a number of alternative hashtags within Twitter (#TOPIC and #TOPIC_ALT), or without an overt marker, or even as part of private messaging between Twitter users, who can then contribute to the public stream again. Moreover, the same topic is usually discussed in parallel across different on-line and off-line media (Facebook, forums, blogs), whose users contribute and process content between the different services. At the same time, the course of the events that relate to or, in fact, constitute the topic develops in reality. This dispersion of topic-related information affects a given isolated hashtag stream: the particular channels generate a pool of possible content to enter the stream. (see Figure 1)

Figure 1. *The complexity of a prototypical topical network evolving on Twitter. All arrows indicate possible channels for content stream.*



Then, agents (*i.e.* users and/or contributors in general) constitute a bridge for these separate channels to feed each other – here the second bottleneck becomes visible. Agents can either process the topical information and then post to the stream of interest, or replicate fragments as well as complete texts within (RT in Figure 1) or across platforms. Thus, we see that content-related bottlenecks interact with agents' biases.

One of such biases is the status of a user. Based on their SNA study, Wu *et al.* (2011) categorize Twitter account holders into "elite" and "ordinary": the former exhibit extremely high followee numbers (up to millions) – the latter have relatively lower audience counts. A user's status determines the visibility of her posts (Wu *et al.* 2011: 706), which means that the elite users generate content accessible to a significantly larger audience. Thus, it is this content that will have greater adaptive power, *i.e.* higher probability for replication (via RT), or simply more significant impact on viewers.

Another bias also relates to the categorization of users, yet under a different criterion. The development of Twitter as well as web technologies in general spawned the so-called *bots*, that is, algorithms programmed to post content automatically across different Internet services. Imagine the unattainable human workload necessary for any of the major news media to post a link to their story across multiple social networks. Therefore, the redundancy of such tasks is handled by bots, which also begin to manage some portion of Twitter accounts. Importantly, the output generated by non-human agents has been reported to exhibit differences in content attraction (Edwards *et al.* 2014).

Perhaps the most obvious bias rests on the assertion that agents, if human, have linguistic systems already entrenched and that this cognitive makeup will influence any evolving sublanguage (Pokornowski and Rogalska 2014). A number of studies conducted indicate that users' behavior differs depending on the language they use in tweets, or their proficiency in it (Hong *et al.* 2011). Similarly, within one language, the choice of orthographic and/or morphological strategies varies depending on a particular variety or dialect (Gouws *et al.* 2011). Extra-linguistic factors, such as users' geo-political context or cultural background can also constitute a major bias. For instance, Chen *et al.* (2013) describe the linguistic strategies used for avoiding censorship by users of the Chinese microblogging platform, Weibo.

Finally, platform-related bottlenecks will encompass all the pressures that stem from Twitter's design as a social networking service. For instance, the microblogging length limit of 140 characters per tweet already makes any Twitter corpus a collection of texts of a very specific genre, which, in turn, explains the stylistic variation found in Twitter stream (see Hu *et al.*

2013). Other pressures relate to specific functionalities, such as the retweet option, enabling either exact (embedded retweet functionality) or partial (via the RT marker) replication of particular content. One important factor affecting the Twitter stream is of course the evolution of the platform itself: new implementations, whether user-driven or provided by the developers, immediately shape information diffusion patterns within the entire network (Liu *et al.* 2014).

The bottlenecks and biases mentioned above need to be accounted for when investigating the evolution of sublanguages emerging around particular topics on Twitter. If carefully controlled, this matrix of pressures and constraints will enable the discovery of crucial adaptive mechanisms that shape the unfolding linguistic code under investigation.

9. Conclusion

The model proposed in this paper constitutes a methodological approach to the study of glossogenetic evolution of sublanguages on the microblogging platform Twitter, or possibly through further modification, on other social networking services. Although related to the Iterated Learning framework, this model has the advantage over the IL approach in its reliance on large masses of authentic language data. For the field of evolutionary linguistics this provides a possibility to arrive at more accurate conclusions concerning the mechanisms that govern language evolution at the glossogenetic level. In turn, a better understanding of the inner dynamics of Twitter content can aid NLP research in the development of processing tools that will be capable of embracing any natural changes occurring within a given stream of data, hence, increasing the accuracy of big data analyses that can have significant implications for our societies.

References

Bengfort, B. (2013). Big Data and Natural Language Processing. http://datacommunitydc.org/blog/2013/05/big-data-and-nlp/. DOA: 15 Aug. 2014.

Broniatowski, D. A., M. J. Paul, and M. Dredze.(2014). Twitter: Big data opportunities. *Inform* 49: 255.

Bruns, A. and J. E. Brugess. (2011). The use of Twitter hashtags in the formation of ad hoc publics. In 6th European Consortium for Political Research General Conference, 25–27 August 2011, University of Iceland, Reykjavik.

Cambria, E. and B. White. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine* 9.2:48–57.

Chen, L.,C. Zhang., and C. Wilson. (2013). Tweeting under pressure: analyzing trending topics and evolving word choice on sina weibo. In *Proceedings of the first ACM conference on Online social networks,* 89–100. ACM.

Chen, M., S. Mao, and Y. Liu. (2014). Big Data: A Survey. *Mobile Networks and Applications* 19.2: 171–209.

Chomsky, N. (2007). Biolinguistic explorations: Design, development, evolution. *International Journal of Philosophical Studies* 15.1: 1–21.

Cogan, P., M. Andrews, , M. Bradonjic, , W. S. Kennedy,, A. Sala, and G. Tucci. (2012). Reconstruction and analysis of twitter conversation graphs. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, 25–31. ACM.

Conte R., N. Gilbert , C. Cioffi-Revilla, G. Deffuant, J. Kertesz, V. Loreto, S. Moat, J.-P. Nadal, A. Sanchez, A. Nowak, A. Flache, M. San Miguel, and D. Helbing. (2012). Manifesto of computational social science. Eur. Phys. J. Special Topics, 214:325–346.

Cunha, E., G. Magno, G. Comarela, V. Almeida, M. A. Gonçalves, and F. Benevenuto. (2011). Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *Proceedings of the Workshop on Languages in Social Media*, 58–65. Association for Computational Linguistics.

Davenport, T. H. and D. J. Patil. (2012). Data Scientist. *Harvard Business Review* 90: 70–76.

Edwards, C., A. Edwards, P. R. Spence, and A. K. Shelton. (2014). Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. *Computers in Human Behavior* 33: 372–376.

Gouws, S., D. Metzler, C. Cai, and E. Hovy. (2011). Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Languages in Social Media*, 20–29. Association for Computational Linguistics.

Hilbert, M. (2014). What Is the Content of the World's Technologically Mediated Information and Communication Capacity: How Much Text, Image, Audio, and Video? *The Information Society* 30.2: 127–143.

Honeycutt, C., and S. C. Herring. (2009). Beyond microblogging: Conversation and collaboration via Twitter. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on,* 1–10. IEEE.

Hong, L., G. Convertino, and E. H. Chi. (2011). Language Matters In Twitter: A Large Scale Study. In *ICWSM*.

Hu, Y., K. Talamadupula, and S. Kambhampati. (2013). Dude, srsly?: The Surprisingly Formal Nature of Twitter's Language. In *ICWSM*.

Hurford, J. R. (1990). Nativist and functional explanations in language acquisition. *Logical issues in language acquisition*, 85–136.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.

Jackendoff, R. (2007a). A whole lot of challenges for linguistics. *Journal of English Linguistics* 35.3: 253–262.

Jackendoff, R. (2007b). Linguistics in cognitive science: The state of the art. *The linguistic review* 24.4: 347–401.

Jagadish, H. V., J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi. (2014). Big data and its technical challenges. *Communications of the ACM*, 57.7: 86–94.

Jasiński, A. Iterated learning of 'evolects' and the dynamics of (re)production of natural language resources. In preparation.

Kirby, S., K. Smith, and H. Brighton. (2004). From UG to universals: Linguistic adaptation through iterated learning. *Studies in Language* 28.3: 587–607.

Kirby, S., H. Cornish, K. and Smith. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105.31: 10681–10686.

Kirby, S. and J. Hurford. (2002). The Emergence of Linguistic Structure: An overview of the Iterated Learning Model, In Cangelosi, A. i Parisi, D. (eds.), Simulating the Evolution of Language. London: Springer Verlag, 121–148.

Kirby, S., T. Griffiths, and K. Smith. (2014). Iterated learning and the evolution of language. *Current opinion in neurobiology* 28: 108–114.

Kwak, H., C. Lee, H. Park, and S. Moon. (2010). What is Twitter, a social network or a news media?" In *Proceedings of the 19th international conference on World wide web*, 591–600. ACM.

Lazer, D., R. Kennedy, G. King, and A. Vespignani. (2014). The Parable of Google Flu: Traps in Big Data Analysis. Science 343.6176 (March 14): 1203–1205.

Letierce, J., A, Passant, J. Breslin, and S. Decker. (2010). Understanding how Twitter is used to spread scientific messages. In Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, April 26–27th, 2010, Raleigh, NC: US.

Liu, Y., C. Kliman-Silver, and A. Mislove. (2014). The Tweets They are a-Changin': Evolution of Twitter Users and Behavior. In *ICWSM*

Macskassy, S. A. (2012). On the Study of Social Interactions in Twitter. In *ICWSM*.

Pokornowski, M. and K. Rogalska. (2014). "Investigating glossogeny via iterated learning methodology: the effect of entrenched linguistic system(s) in human agents". Presented at the Languages in Contact 2014, (17 May 2014) Wroclaw, Poland.

Provost, F., and T. Fawcett. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data* 1.1: 51–59.

Rossi, L., and M. Magnani. (2012). Conversation Practices and Network Structure in Twitter. In *ICWSM*.

Wacewicz, S. (2013). Ewolucja języka – współczesne kontrowersje. In Stalmaszczyk, P. (ed.) *Metodologie językoznawstwa. 1. Ewolucja języka. Ewolucja teorii językoznawczych*. [Evolution of Language – contemporary controversies In Stalmaszczyk P. (ed.) *Linguistic methodologies: 1. Evolution of Language.*

The Evolution of Linguistic Theories]. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, 11–26.

Weller., K., E. Dröge, and C. Puschmann. (2011). Citation Analysis in Twitter: Approaches for Defining and Measuring Information Flows within Tweets during Scientific Conferences. In *# MSM,* 1–12.

Wu, S., J. M. Hofman, W. A. Mason, and D. J. Watts. (2011). Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web,* 705–714 . ACM.

Zappavigna, M. (2011). Visualizing logogenesis: Preserving the dynamics of meaning. *Semiotic Margins: Meaning in Multimodalites*. London: Continuum, 211–228.