A Love-Hate Relationship for Big Data and Linguistics: Present Issues and Future Possibilities

Giorgio Maria Di Nunzio
Department of Information Engineering
University of Padua
Via Gradenigo 6/a, 35131 Padua, Italy
giorgiomaria.dinunzio@unipd.it

Cecilia Poletto
Department of Linguistic and Literary Studies
University of Padua
Piazzetta Folena 1, 35137 Padua, Italy
cecilia.poletto@unipd.it

ABSTRACT

In this paper, we present an overview of some issues related to the use of Big Data in the area of Linguistics that have been debated in workshops and conferences in the last two years. We also consider some requirements that "big" linguistic databases should have in order to tackle some of these issues; finally, we discuss a set of possible interactive visualization approaches of large datasets that may have an impact in this research field.

Keywords

Linguistic Databases, Geolinguistics, Linguistic Maps.

1. INTRODUCTION

The availability of large text collections for the research field of Linguistics has increased significantly in the last years. Researchers have turned their attention to large linguistic databases and corpora of all kinds, to experimental results, and to many other types of data sources. This development went hand in hand with an expansion of the scope of the theories, and collaborations with e.g. historical linguistics, dialectologists, sociolinguists, and psycholinguists. [1] These linguistic databases open new opportunities for linguistic research, but they may be problematic in terms of representativeness and accuracy of the data. Although the term "big data" is well defined in several empirical domains, in comparative linguistics the discussion on the size and properties big data should have in order to be considered as such has started only very recently. At present, it is the object of ongoing discussion and runs in parallel with the construction of comparative linguistic data bases, i.e. of essential tools to large empirical investigations. Even inside linguistics, the various domains are not homogenous, since phonology has started the discussion on big data and gathering data on the phonological inventory of many languages much earlier than other domains, like syntax.

In the last two years, the debate about the impact of Big Data in this research field has shown two different points of view: on the one hand, we have researchers who are aware of the challenges that the creation and use of big data for linguistics poses, but they are mainly positive about it given the benefits that these datasets can provide, i.e. new kinds of evidence about pragmatic, sociolinguistic and even syntactic aspects of linguistic events.[2] For example, professor Liberman observes that[3]

> "we can now study linguistic patterns in space, time, and cultural context, on a scale three to six orders of magnitude greater than in the past, and simultaneously in much greater detail than before. [...] Of course, our observations may not be correct or general, because they depend on counting things in specific datasets with specific characteristics. But the same problem exists even more seriously for the answers we get from any other methods. And as long as we have data from a variety of different settings [...] it is easy to check the generality of our results."

The idea is to use these mega-corpora with judgement because they can be crucial in understanding languages and language variation. It is important to understand whether and how these data correlate with data from more carefully constructed, balanced corpora [10]. In 2013, an interesting discussion about the Global Lexicostatistical Database (GLD)[4] took place in a conference at Max-Planck-Institut about "Historicizing Big Data".[5] In particular, [13] presented this big data stage as a continuity rather than a rupture in the research field. She showed that theory is also built into the database infrastructure of contemporary linguistics research in the GLD and, while this collaborative online database is new, it brings together two century-old, formerly competing traditions in linguistics. One year after, in 2014, a panel of the Joint British Academy and Philological Society[6] discussed some fundamental questions about how the results of traditional scholarship can be integrated with those derived by digital methods as well as how we can

[1] http://www.meertens.knaw.nl/baddata/?page_id=5

[2] https://www.helsinki.fi/en/researchgroups/varieng/d2e-from-data-to-evidence

[3] http://www.theguardian.com/education/2014/may/07/what-big-data-tells-about-language

[4] The major goal of this project is to put together and make available, for specialists and the general public alike, the most complete and thoroughly annotated collection of basic wordlists of the world's language

[5] https://www.mpiwg-berlin.mpg.de/en/research/projects/deptii_kaplan_reconstruction

[6] http://www.britac.ac.uk/events/2014/Language_Linguistics_Data_Explosion.cfm

In Proceedings of 1st International Workshop on Accessing Cultural Heritage at Scale (ACHS'16), June 22, 2016, Newark, NJ, USA.

measure the impact of such challenges on diverse areas of language-study.

On the other hand, we have other researchers who have a critical positions towards the use of Big Data. In 2016, the main focus of a workshop at the Meertens Institute was about dealing with bad data in linguistic theory.[7] The participants debated about different classes of problems that big data may have on linguistics research, such as incomplete data, noisy data, one-sided data, and conflicting data. For example, [9] showed that very large dataset are potentially very useful to improve our understanding of some linguistic theoretical analysis, but at the same time there are important reasons to consider that are in part complementary to the ones already mentioned, i.e. missing data, different methodologies, different frequencies, and different systems. Therefore, it becomes crucial to understand to what extent which of these properties has an impact on any theoretical conclusions that can be drawn from this data set.

2. BIG GEOLINGUISTICS DATABASES

Research in language variation allows linguists to understand the fundamental principles that underlie language systems and grammatical changes in time and space. Geolinguistics is an interdisciplinary field that aims at mapping the geographical distribution of phenomena which are mainly due to processes of grammatical principled changes [12]. In this context, the linguistic atlas has proved to be a vital tool and product of geolinguistics since the earliest stages of the field, and it has provided a stage for the incorporation of modern GIS. In the last two decades, several large-scale databases of linguistic material of various types have been developed worldwide [7]. One of the basic problems we have to deal with geolinguistic databases data is related to the qualitatively and quantitatively different types of data that have to be classified and retrieved [4]. A geolinguistic database with the function of the traditional linguistic atlases contains a variety of data that are beyond the simple linguistic information (for example, geographic locations, the type of inquiries adopted to gather the data, the speakers who have delivered the data, and so on), all of them being relevant to the geolinguistic analysis.

The interaction between linguistic and geographical information becomes crucial in situations when a given linguistic phenomenon is found in the same geographical area as another, or the two linguistic phenomena are in geographically disjunct areas, or the area of the first implies the area of the second. The visualization of the geographic distribution of these phenomena represents precious information for the linguist and should be immediately retrievable from the interface. Another important facet for geolinguistic research concerns the test subjects used to gather the data on the field. They might provide input to investigate how language changes over time in a given geographical space. Since the data to be combined are of different origin and are to be classified according to different parameters, a careful planning of the structure of the database as well as the interactive visualization is necessary to develop a system which has the properties of durability and wide usage among researchers that justify such an expensive enterprise.

2.1 New insights through linguistic maps

The problem concerning the tools to filter incomplete, noisy or more generally "bad data" clearly depends on the type of research aims. If the aim is a qualitative and not a quantitative investigation, we definitely need a new set of procedures to compare big amounts of data. The degree of fine-grained distinctions necessary in a qualitative enterprise is clearly much higher than the one generally required in quantitative research and which cannot be provided by standard statistical approaches already used in language acquisition and psycho-linguistics.

Big linguistic enterprises, like data bases, atlases and all sorts of corpora, always contain a certain amount of "noise". They are by definition always both incomplete and inaccurate when the linguistic hypothesis we want to test is already very detailed and precise. On the other hand, data are also incomplete even when we adopt a qualitative procedure which investigates a smaller subset of data. For instance, Buchstaller and Corrigan in[8] show that the results we can obtain also depend of the type of task the test subjects have been asked to perform, and that the best policy is not only to control for all possible factors intervening in the experiment we are performing, but also to combine different tasks to single out stable linguistic generalizations that are not prone to be simple task effects. This means that the amount of data we are considering is irrelevant with respect to the problem of how complete and reliable our data set can be. Since data are always inevitably incomplete, what we have to do is develop new strategies to compensate for the inaccuracy and incompleteness of the data. In this respect, it can be useful to consider some strategies that can help us to find out interesting theoretical clues even in data that provide by definition a coarse-grained picture of the linguistic reality. If it is true that big data is never precise enough for a very detailed hypothesis, we can still try to exploit the peculiarity of a blurred but very big image to single out the general outlines of the linguistic panorama, which would remain otherwise uncovered. In this way, using big data mining can nicely complement our introspective type of empirical evidence in the spirit of Buchstaller and Corrigan, who suggest the combination of different test strategies. The reason why big data are always too "noisy" is that we do not treat them in the right way, i.e. the questions we ask are not adapted to the type of evidence we have. The unavoidable conclusion is that new strategies to represent and exploit the data we have at our disposal have to be developed. The general gist of the solution to the problem we will present is the following: up to now we have only used big corpora to look at the presence versus absence of a given phenomenon X in a given language L and related it to other phenomena.

An innovative way to think about big data and tailor our questions on the linguistic evidence provided by big data is to consider the type of geographical variation itself as a clue indicating different natural classes of linguistic phenomena. It is possible to single out at least three distributional patterns and determine to which type of phenomenon each type of variation is related.

2.1.1 "Classic method"

The first "classic" method to be taken into account and further developed with new technical representation tools is the one adopted by geolinguists since the beginning of the discipline, i.e. the one of comparing the geographical distribution of different phenomena inside a genetically ho-

[7] ttp://www.meertens.knaw.nl/baddata/

mogeneous linguistic area and consider the theoretical import of different distributional patterns. In particular there are three clearly identifiable distributional patterns that can provide us with new insights into the linguistic system when we consider the distribution of two phenomena. The two phenomena can a) completely overlap, b) be in complementary distribution or c) one can be included into the other on a linguistic map representing both of them. When they completely overlap, this can be interpreted as having the two phenomena depend on a single abstract property. When they are in complementary distribution, this means that they occupy the same space inside the linguistic system, i.e. they satisfy the same requirement. Complementary distribution thus means that two phenomena are alternative checkers of the same linguistic property, so that they exclude each other. When one is included in the other, this can be taken as an indication that the wider phenomenon partially depends on a setting that also the smaller one shares, but has additional requirements. In other terms, this could mean that the phenomenon that is more largely represented is a necessary but not sufficient condition for the occurrence of the second. Hence, the pure geographical distribution of co-varying phenomena can provide us with interesting clues to interpret the data.

Although it was never really formalized, this type of methodology has been used by traditional dialectologists and is still used today in formal frameworks and can be only be adopted when we are comparing two phenomena and trying to establish whether they are intrinsically related or not. Still, the distribution we find could only be by chance, but if we have enough languages, the probability that we only have to do with chance reduces progressively the bigger our sample is. Which means, big data are a valuable source to linguistic investigation, but they have better be really big.

2.1.2 *"Leopard spots"*

Another type of geographical distribution which can provide us with interesting observations that we would not be able to see on the basis of a detailed qualitative investigation or even analyzing big data on the basis of other devices that are not as visually immediately interpretable as maps are. This type of distribution is called by traditional dialectologists "leopard spots" because the phenomenon under study occurs precisely with an apparently random distribution which however covers the whole area taken into account. This type of distribution is generally found when we deal with a phenomenon that is only possible when a specific complex constellation of factors is instantiated in the same language. Since the phenomenon depends on several factors that do not depend on each other in any sense (either implication of exclusion), we find it only where all the factors cluster together and this can happen in various points of the area. The study of this type of variation can lead us to find out exactly what the complex prerequisites are that lead to the occurrence of the phenomenon under study. This special type of distribution has been discovered in all areas that include strictly genetically related languages, while it is not found on linguistic maps treating languages which are very different, for instance in linguistic typology. Since it is typical of microvariation and not macrovariation, it constitutes a very powerful tool to pin down phenomena that depend on complex clusters of often unrelated properties.

2.1.3 *"Genetically" related languages*

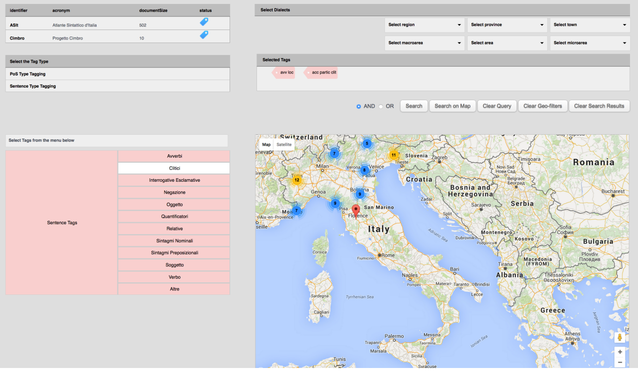
A third type of variational data that can be exploited once it is set on a map, has to do with a very simple observation and has never been used up to now, although it is actually very simple. It is possible to extract syntactic and semantic observations from lexical data simply by looking at the type and number of possible lexical forms for the same element starting from a simple but rather strong hypothesis: the index of lexical variation of a functional word like auxiliaries, prepositions, pronouns etc. within a genetically related set of languages co-varies with the semantic and syntactic complexity of the item itself. Therefore, a simple count of the possible forms across the area considered gives us information about the semantic/syntactic complexity of the item we are investigating. This evidently only works within a domain where all languages are genetically related, i.e. where the original etymological set of possible elements is constrained by the fact that all languages considered come from the same source language (like for instance all Romance languages share their major lexical endowment which comes from Latin). This means that a rather simple count of the possible lexical etymological sources used in a set of genetically related languages gives us very precise indications of the featural primitive components the functional element is made up of. This is a tool that has never been tried out up to now precisely because no one has ever thought of using the massive amount of data we now have at our disposal in this way.

3. DISCUSSION

This new way to think about big data and linguistics requires some thoughts at different levels: the choice of the type of visual interaction; an efficient data structure to store, organize and retrieve linguistic data; an evaluation of the implemented system. The exploration of new visualization and interaction systems with a geographic map were presented in the ASIt (Syntactic Atlas of Italy) project [3, 1, 2]. In Figure 1 we show an interactive interface used to search a linguistic database with sentence tags or POS tags. The results of this map can be explored and studied by a linguist in the way described in Section 2.1.1. One of our proposals in this paper is to extend the 'classic' view of linguistic data described in Section 2.1.1 to more complex interactions with the map, like overlapping two or more maps with different level of transparency and, for example, highlights the 'leopard spots' described in Section 2.1.2.

The second point is the study of the 'right' model and data structure for large linguistic datasets. On the one hand, we have the problem of designing systems that should give access to digital objects that may be stored in different institutions, i.e archives and museums; therefore, the interoperability among the Digital Library System which manage the digital resources of these institutions is a key concern [5]. For this purpose, the working group of open data in linguistics has recently promoted the idea and definition of open data in Linguistics and in particular to the use of Linked Open Data (LOD) to implement it. The LOD paradigm refers to a set of best practices for publishing data on the Web[8] and it is based on a standardized data model. In the ASIt project, we have proposed a LOD approach for increasing the level of interoperability of geolinguistic applications and the reuse

[8] http://www.w3.org/DesignIssues/LinkedData.html



identifier	acronym	documentSize	status
ASIt	Atlante Sintattico d'Italia	502	
Cimbro	Progetto Cimbro	10	

Figure 1: Interactive maps of the ASIt project.

of the data. In particular, we defined an extensible ontology for geolinguistic resources based on the common ground defined by current European linguistic projects and we applied this ontology on top on a real linguistic dataset [11, 7]. Nevertheless, a study of the efficiency of the LOD approach on very large dataset is still to be completed.

Last but not least, since the system we are trying to develop requires interaction, interoperability and efficiency, we need a validation of the system. In the CULTURA project, for example, the IPSA system was evaluated by both expert researchers and students in order to collect all the different points of view of the users of a digital library, in its transition from an isolated archive to an archive fully immersed in a new adaptive environment [6].

4. CONCLUSIONS

The traditional geolinguistic tool of linguistic maps can provide new and important indications to theoretically interpret linguistic phenomena. We described the fundamental requirements that are needed to adapt classical linguistic maps in order to carry out more sophisticated analyses.

This is only possible when two conditions are met: the set of languages investigated is genetically related and we are really dealing with big data. These methods can also compensate with the inaccuracy of the data, since they do not need to be very detailed for us to gather an idea of the type of distributional pattern we are dealing with. Evidently, this type of methodology is not intended as a substitution of more traditional methods, but complements other research methodologies in an integrated view of research.

Acknowledgments

We are grateful to Maristella Agosti for the fruitful discussions and her useful comments on an earlier version of this paper.

5. REFERENCES

[1] M. Agosti, B. Alber, P. Benincà, G. M. Di Nunzio, M. Dussin, R. Miotto, D. Pescarini, S. Rabanus, and A. Tomaselli. Asit: A grammatical survey of italian dialects and cimbrian: Fieldwork, data management, and linguistic analysis. In *Digital Libraries and Archives - 7th Italian Research Conference, IRCDL 2011, Pisa, Italy, January 20-21, 2011. Revised Papers*, pages 100–103, 2011.

[2] M. Agosti, B. Alber, G. M. Di Nunzio, M. Dussin, S. Rabanus, and A. Tomaselli. A curated database for linguistic research: The test case of cimbrian varieties. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 2230–2236, 2012.

[3] M. Agosti, P. Benincà, G. M. Di Nunzio, R. Miotto, and D. Pescarini. A digital library effort to support the building of grammatical resources for italian dialects. In *Digital Libraries - 6th Italian Research Conference, IRCDL 2010, Padua, Italy, January 28-29, 2010. Revised Selected Papers*, pages 89–100, 2010.

[4] M. Agosti, E. Di Buccio, G. M. Di Nunzio, C. Poletto, and E. Rinke. Designing A Long Lasting Linguistic Project: The Case Study of ASIt. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016), Portoroz, Slovenia, May 23-28, 2016.*, page In press., 2016.

[5] M. Agosti, N. Ferro, and G. Silvello. Digital library interoperability at high level of abstraction. *Future Generation Computer Systems*, 55:129–146, 2 2016.

[6] M. Agosti, M. Manfioletti, N. Orio, C. Ponchia, and G. Silvello. *Bridging Between Cultural Heritage Institutions: 9th Italian Research Conference, IRCDL 2013, Rome, Italy, January 31–February 1, 2013, Revised Selected Papers*, chapter The Evaluation Approach of IPSA@CULTURA, pages 147–152. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[7] E. D. Buccio, G. M. Di Nunzio, and G. Silvello. A curated and evolving linguistic linked dataset. *Semantic Web*, 4(3):265–270, 2013.

[8] I. Buchstaller and K. Corrigan. Making intuitions work: Testing instruments for measuring dialect syntax. In W. Maguire and A. McMahon, editors, *Analysing Variation in English*, pages 30–48. Cambridge University Press, 2011.

[9] J. V. Craenenbroeck. Handle your verb clusters with care. Dealing with bad data in linguistic theory. Amsterdam, the Netherlands., March 2016.

[10] M. Davies. Why size alone is not enough: the importance of historical, genre-based, and dialectal variation in language. D2E Conference, "From data to evidence in English language research: Big data, rich data, uncharted data", Helsinki, October 2015.

[11] E. Di Buccio, G. M. Di Nunzio, and G. Silvello. A linked open data approach for geolinguistics applications. *IJMSO*, 9(1):29–41, 2014.

[12] S. Hoch and J. J. Hayes. Geolinguistics: The Incorporation of Geographic Information Systems and Science. *The Geographical Bulletin*, 51(1):23–36, 2010.

[13] J. Kaplan. The Global Lexicostatistical Database: Integrating Traditions in Long-Range Historical Linguistics. "Historicizing Big Data" Conference, Max-Planck-Institut, October 31 – November 2, 2013, October 2013.