Reliability and Meta-reliability of Language Resources: Ready to initiate the Integrity Debate?

António Branco

University of Lisbon
Antonio.Branco@di.fc.ul.pt

1 Introduction

Given the increasing complexity and expertise involved in the development of language resources, there have been a growing interest in finding mechanisms so that the designing and the development of language resources may be taken as a first class citizen in terms of scientific work, and accordingly cvs and careers of individual researchers can be fairly credited and rewarded for that. Ongoing initiatives such as the international standard language resource number (Choukri, 2013), or the studies on metrics to ascertain the reliability of linguistically interpreted data sets (e.g. Artstein and Poesio, 2008) are just a few illustrative examples of this trend.

Concomitant to this movement of reinforced scientific credibility, but in an opposite direction, there have been appearing worrying signs that, in what concerns mature and well established scientific fields, scientific activities and results may be untrustable to an extent larger than possibly expected and acceptable. That this issue has recently hit the mass media[1] is but an indicator of the volume and relevance of these signs, whose assessment and discussion became unavoidable across all sectors of the international scientific system.

These signs have been related, for instance, to the realization that for a considerable proportion of published results their replication is not being obtained by independent researchers (e.g. Florian et al., 2011; Belgley and Ellis, 2012); to the deliberately falsified submissions of papers for publication, with fabricated errors and fake authors, which get easily accepted even in respectable journals (Bohannon, 2013); or to the outcome of inquiries to scientists on questionable practices, with scores higher than one might expect

[1] Unreliable Research: Trouble at the Lab, *The Economist*, October 19th, 2013.
Hiltzik, Michael, 2013, Science has Lost its Way, at a big Cost to Humanity, *Los Angeles Times*, October 17, 2013.
Zimmer, Carl, 2012, A Sharp Rise in Retractions Prompts Calls for Reform, *The New York Times,* April 16, 2012
Begley, Sharon, 2012, In Cancer Science, Many "Discoveries" don't Hold up, *Reuters*, March 28th, 2012.
Nail, Gautam, 2011, Scientists' Elusive Goal: Reproducing Study Results, *The Wall Street Journal*, December 2, 2011.

or would be ready to accept (Fanelli, 2009). For a receent and updated overview and further references on these signs, see (Stodden, 2013).

A number of causes have been aired for these state of affairs including, among others, increasingly sloppy reviewing; the growing number of so-called "minimal-threshold" journals; policies for publication that do not require the sharing of at least the raw or primary data; or the non disclosure of the software developed and used to obtain the results published. These causes have deserved serious scrutiny, including in the "World Conference on Research Integrity", whose third edition was held this year.[2]

Underneath these immediate causes, a number of factors have been pointed out, including, for instance, not enough negative incentives or peer-pressure to hamper the above practices; career and promotion pressure too biased for quantity; widespread disinterest on negative results as an intrinsic part of the scientific progress; widespread disfavoring of activities of replication by funding agencies; poor or non existent retraction procedures for results that are eventually noticed to be wrong or flawed after having been published; ideological pressure to get immediate financial return from research results; etc.

In Bill Frezza's bold opinion, the financial pressure on the scientific system "has created a moral hazard to scientific integrity no less threatening than the moral hazard to financial integrity that recently destroyed our banking system." (Frezza, 2011).

In the present invited talk at the 12th Workshop on Treebanks and Linguistic Theory, I am interested in contributing to initiate a debate on what part of the above issues may be recognized as having the conditions to be eventually happening also in our field, what part does not apply to it given its specific nature, and what may be the risks that may be specific to it. The ultimate goal of this exercise is to contribute for the reinforcement of the scientific credibility of language resources, and to the integrity of our scientific work around them.

Before proceeding, a word of clarification is in order, in particular to indicate what this talk is not about. It is not about what one might term as issues of empirical adequacy of linguistically interpreted data sets. These are issues related to the adequate interpretation of the markables. For instance, issues that occur if in the annotation of a corpus, as a result of a flawed design, the annotation principles or guidelines would wrongly require that what are standard grammatical prepositions be mistakenly annotated as adjectives, etc. These are the issues addressed, for instance, in (Zaenen, 2006).

It is not about issues of reliability of annotated data sets either. These are issues that are related to the adequate definition of the annotation methodology in view of minimizing errors in the application of the annotation guidelines, and that can be monitored by metrics involving inter-

[2] http://www.wcri2013.org

annotator agreement, etc. These are the issues addressed, for instance, in (Artstein and Paoesion, 2008).

This talk is about integrity issues that are associated to the overall scientific ecosystem where the development of language resources and the research around it takes place. These are issues related to the overall conditions that support the credibility of and trust on the scientific work and its results, and that remain to be addressed even if the issues on empirical adequacy and reliability of the data sets are eventually settled.

2 Essential replication

Let us take two key processes contributing for the integrity of scientific activity and results, reviewing and replication, the former applying before (or leading to) the publication of results, and the latter applying after these have been published. The point worth noting at this juncture is that the need for replication does not result from the fact the ideal of flawless reviewing cannot be attained. Replication is a first class citizen here and it is still needed to play a crucial role even in case flawless reviewing could have been ensured.

For the sake of concreteness, let us consider the example of natural sciences. And to keep the reflection at a general enough level, let us understand that to a large extent, advancements and discoveries reported in papers are obtained as a result of new ways on how to gather primary data and/or how to analyze them into secondary data and empirically supported generalizations. If an ideal flawless review process could be possible, for a top-scoring paper accepted for publication, in essence the reviewers would then have said ok to what? In essence, to what one would call, for lack of a better and more encompassing term, the "methodological" aspects reported in the paper.

For the sake of the point, let us leave intentional misconduct or fabrication of data aside. In spite of the existence of a correct methodology to collect the primary data, their actual gathering may have gone wrong as a consequence of some clerical error or some inadvertent practical slips. Likewise, the analysis into secondary data and generalizations may have not been appropriately executed also due to some fortuitous reasons. The point here is that, in general, for rich and complex enough data, reviewers have no means to detect this kind of problems, unless they would also run the experiments themselves and executed the analysis of the data. But that is what replication is all about, and for obvious practical reasons, it is not and cannot be under the scope of the reviewing process.

3 Emerging validation

How can these considerations be transposed to or help to think about the field of language resources, where the ultimate goal is the development of primary data (to be used at subsequent technological and scientific activities)? At a general enough level of appreciation, we should then start with the note that a typical research paper in our area focus in the first of the two parts indicated above. Though it tends to be seen as a positive feature that papers may report also on technological solutions or tools supported by the data set whose development is being reported, the key topic is clearly which methodological novelties are involved (e.g. a new bootstrapping approach, new languages involved, new relations between previously available data, etc.) and which data set, i.e. primary data, that these innovations have eventually led to.

Let us transpose the question above to our area: If an ideal flawless review process was possible, for a top-scoring paper accepted for publication, in essence the reviewers would then have said ok to what? Again, in essence, to the methodological aspects reported in the paper. And again, in spite of the existence of a correct methodology to collect the primary data, their actual gathering may have inadvertently gone wrong for a number of fortuitous reasons. And that is where and why "replication" has its key role to play.

And here we arrive at an important point of our discussion: what exactly can be, or should be understood as replication, or as playing the role of replication, in this research area of language resources?

Replication permits to go beyond the mere verification of the methodological issues by reviewers, as these are reported in successfully published papers. It permits to check if the execution of those methodological steps, procedures, calculations, processes, etc. actually lead to the results that are being reported. For the area of language resources, in a very narrow and strict sense, this might translate into redoing the data set whose development is reported in a given paper, which clearly is completely out of question for obvious practical reasons. In a less narrow and more sensible sense, this may translate into checking, even if only by an as smart sampling as possible, whether the data set that resulted is actually the one being announced in the paper.

As replication is different of and out of the scope of reviewing in natural sciences, also here in our area whatever the details of this validation process may be, it is not an assignment for reviewers. For one, because for a large number of papers on languages resources, the data sets whose development is being reported are not publicly made available by their authors. But even if they were, and in the growing number of those resources that are actually made available at the moment of the publication of the respective papers, it is obvious that reviewers have no practical conditions to proceed with such validation, which in the case of language resources may play the role analogous to the one replication plays in other areas.

As replication of experiments is a key element in the integrity of the scientific ecosystem of other sciences, validation of language resources cannot but be a key element for the integrity of scientific activities in our area.

4 Illusory impactfulness

It may be tempting to consider that in the case of language resources, their validation is eventually taken care of not at a specific moment or in some dedicated occasion or explicit procedure, but that this just happens implicitly by the "invisible hand" of the different impact of the different resources in the community of researchers and users. A given resource has a larger impact if it is used more frequently and referred to more often in a larger number of papers. But the level of impact of a resource illusoriously correlates with the possible level at which such resource had been validated, even if supposedly by the mere effect of the usage that the community is doing of it.

For languages for which there is a small community of researchers working on it, and little or no funding exists to do so, a resource referred to only a very few times may be a perfectly developed data set, in accordance to the respective methodological principles and guidelines, that may happen to be fully adequate in linguistic terms. The same holds for resources that support work on less researched topics, which comparatively may receive a very small number of references and yet be an extraordinarily well-developed resource, which would top score in any rigorous validation process.

In the opposite direction, it occurs also that a resource may have a widespread usage and receive a high number of references and yet its validation would indicate suboptimal scores (Van Halteren, 2000; Eskin, 2000; Dickinson and Meurers, 2003; Tylman and Simov, 2004; Dickinson and Meurers, 2005). It is enough, for instance, that it is the first of its kind for English and/or supports research in a very hot topic.

Current mainstream research on natural language processing is about getting increasingly better evaluation scores for the relevant type of tools or applications while working with some given data sets (to ensure comparability), which *ipso facto* become the de facto standard data sets. And this can be pursued, and is actually pursued, whether or not those data sets had been correctly developed or had been gone through any validation process.

As Annie Zaenen put it in a humorous way when discussing the specific case of the development of language resources annotated for coreference: "Of course, as long as the task is to provide material to develop and refine machine-learning techniques, much of this doesn't matter. Whether *Henry Higgins* and *Eliza Doolittle* are referring to the same entity or not is of no interest in that context. The technique has only to show that if it is told that they are coreferent because they had the same job (even at different

moments), then it can also learn that George Bush and his father are coreferent." (Zaenen, 2006, p.579).

5 Putting on the agenda

For other long-established scientific areas, the discussion on replication of experiments and other integrity aspects of the scientific work has definitely made its way into the public agenda on science. And the discussion on mechanisms, conditions and incentives to foster, support, fund or perform replication has arrived to stay.[3] By the same token, we should bring the discussion on the validation of resources to the agenda of our community, and add it to other possible forward-looking issues currently aimed at strengthening the conditions of our research work.

When considering the actual enormous amount of effort, time and perseverance that is necessary to put in place a large enough data set that may be annotated with some quite sophisticated linguistic interpretation, under some stringent reliability ensuring methodology, one has to admit that the effort and conditions needed to publish a paper reporting on its development or fill in its metadata record, and get credited for it, is incomparably much ligther. Validation is a crucial element to help preventing and diverting possibly unduly inflated or even void reporting.

The organizations, initiatives or platforms operating the distribution of language resources, such as ELDA, LDC, OLAC, META-SHARE or CLARIN among others, have been driving forces of a continuous endeavor to support and foster the research area of language resources. In my view, it is only natural that, in order for them to evolve along the natural progression of the field and its new demands, the community of researchers expects that the mission of these organizations be extended. In particular, we can expect that these organizations play a key role in contributing to research integrity by being independent stakeholders to whom the role of validating language resources is trusted.

It is certain that in their regular daily operation, the language resources distribution organizations have been proceeding with instrumental verification of the resources that they receive to be distributed, at least to check whether the content of the packages match the description of the

[3] The Reproducibility Initiative (www.scienceexchange.com/reproducibility) was launched in 2012 by several prominent scientific journals and organizations in response to revelations from the pharmaceutical industry that a large proportion of published cancer research cannot be reproduced. It intends to identify and reward high quality reproducible research through independent validation of key experimental results.
The Center for Open Science (centerforopenscience.org) is a non-profit organization founded in January 2013 to increase openness, integrity, and reproducibility of scientific research.

resource provided in its metadata record.[4] But given the discussion above, the point is that this may need to be re-addressed under an entirely new light, and under renewed conditions. More than being just an internal procedure, validation of language resources plays a unique role in the whole ecosystem where the research work on and around language may strive, raises its profile, and hope to keep progressing according to the highest scientific standards.

How the distribution organizations may fulfill this role, assume this responsibility and make a key contribution for the progress of the area is a much-needed debate, which should welcome different views from different actors, and which the present paper would like to trigger. I would though dare to venture that at least a couple of ingredients will be crucial: the validation of language resources should be systematic and public.

For different types of datasets, practically feasible and *de facto* standard procedures should emerge on how to proceed with their validation.

And, as a way of a badge of validity, the metadata record of each resource should publicly indicate which kind of validation procedure it was submitted to, and what were the scores obtained for the different validation parameters if applicable.

Clearly, this will bring the language resources distribution organizations from the level of being instrumental supporters to be key players in the sustainability of the whole area, representing and ensuring a much needed self-regulatory endeavor of the scientific community they were aimed to serve when they were initially set up.

Acknowledgements

This research has received partial funding from the EC's FP7 (FP7/2007-2013) under grant agreement number 610516: "QTLeap: Quality Translation by Deep Language Engineering Approaches".

References

Unreliable Research: Trouble at the Lab, *The Economist*, October 19, 2013, online edition. http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble

Artstein, Ron, and Maximo Poesio, 2008, Inter-Coder Agreement for Computational Linguistics, *Computational Linguistics*, 34(4), pp.555-596.

[4] http://catalogue.elra.info
https://www.ldc.upenn.edu/data-management/using
(consulted on November 25, 2013)

Begley, Sharon, 2012, In cancer science, many "discoveries" don't hold up, *Reuters*, March 28th, 2012, online edition. http://www.reuters.com/article/2012/03/28/us-science-cancer-idUSBRE82R12P20120328

Begley, Glenn and Lee Ellis, 2012, Drug development: Raise standards for preclinical cancer research, *Nature*, 483, pp.531-533.

Bohannon, John, 2013, Who's Afraid of Peer Review?, *Science*, 342, pp.60-65.

Choukri, Khalid, 2013, International Standard Language Resource Number, Presentation at the *ELRA International Workshop on Sharing Language Resources: Landscape and Trends*, Paris, November 19-20.

Dickinson, Markus and W. Detmar Meurers, 2003a, Detecting Errors in Part-of-Speech Annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*. Budapest, Hungary, pp. 107–114.

Dickinson, Markus and W. Detmar Meurers, 2003b, Detecting Inconsistencies in Treebanks. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*. Va¨xjo¨, Sweden.

Dickinson, Markus and W. Detmar Meurers, 2005, Towards Detecting Annotation Errors in Spoken Language Corpora, In *Proceedings of the 15th Nordic Conference of Computational Linguistics (NODALIDA 2005)*, Special Session on Treebanks for Spoken Language and Discourse.

Eskin, Eleazar, 2000, Automatic Corpus Correction with Anomaly Detection. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-00)*. Seattle, Washington.

Fanelli, Daniele, 2009 How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data, *PLOS ONE*, DOI: 10.1371/journal.pone.0005738.

Frezza, Bill, 2011, The Financially Driven Erosion of Scientific Integrity, *Real Clear Markets*, December 5, 3011, online edition. http://www.realclearmarkets.com/articles/2011/12/05/the_financially_driven_erosion_of_scientific_integrity_99401.html.

Hiltzik, Michael, 2013, Science has Lost its Way, at a big Cost to Humanity, *Los Angeles Times*, October 17, 2013, online edition, http://www.latimes.com/business/la-fi-hiltzik-20131027,0,1228881.column#ixzz2lT8zjZWD

Nail, Gautam, 2011, Scientists' Elusive Goal: Reproducing Study Results, *The Wall Street Journal*, December 2, 2011, online edition, http://online.wsj.com/news/articles/SB10001424052970203764804577059841672541590

Prinz, Florian, Thomas Schlange and Khusru Asadullah, 2011, Believe it or not: how much can we rely on published data on potential drug targets?, *Nature Reviews Drug Discovery* 10, 712.

Stodden, Victoria, 2013, Resolving Irreproducibility in Empirical and Computational Research, *IMS Bulletin Online*, American Institute of Mathematical Statistics. http://bulletin.imstat.org/2013/11/resolving-irreproducibility-in-empirical-and-computational-research/

Ule, Tylman and Kiril Simov, 2004, Unexpected Productions May Well be Errors. In *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal.

van Halteren, Hans, 2000, The Detection of Inconsistency in Manually Tagged Text. In Anne Abeilleé, Thorsten Brants and Hans Uszkoreit (eds.), *Proceedings of the Second Workshop on Linguistically Interpreted Corpora (LINC-00)*, Luxembourg.

Zaenen, Annie, 2006, Mark-up Barking Up the Wrong Tree, *Computational Linguistics*, 32 (4), pp. 577-580.

Zimmer, Carl, 2012, A Sharp Rise in Retractions Prompts Calls for Reform, *The New York Times*, April 16, 2012, online version, http://www.nytimes.com/2012/04/17/science/rise-in-scientific-journal-retractions-prompts-calls-for-reform.html?_r=0