We Are Depleting Our Research Subject as We Are Investigating It: In Language Technology, more Replication and Diversity Are Needed

António Branco
University of Lisbon
NLX-Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências
Campo Grande, 1749-016 Lisboa, Portugal
antonio.branco@di.fc.ul.pt

Abstract

In this paper, we present an analysis indicating that, in language technology, as we are investigating natural language we are contributing to deplete it in the sense that we are contributing to reduce the diversity of languages. To address this circumstance, we propose that more replication and reproduction and more language diversity need to be taken into account in our research activities.

Keywords: science ethics, science policy, language technology, human language diversity, replication of research results, language policy.

1. Introduction

Natural language is a most extraordinary object of scientific inquiry lending itself to be researched at least as a referential symbolic system, a socially effective type of behavior or a class of specialized mental activities, and hopefully one day as a principled unified combination of all its dimensions. As it is at the core of what distinctively human nature may be, the approximately 7 000 human languages existing in our planet are a most valuable treasure trove for scientific inquiry on the human brain, mind and behavior, and for advancing our understanding of ourselves and finding better technological solutions that improve our life and heal us.

While informed laypersons are aware of the dramatic consequences of the depletion of important natural resources, from energy to bio-diversity and including potable water, ozone layer among several others, they are much less, or not at all aware of the threat hanging over language diversity. Around one third of the world languages are at present vulnerable to become extinct according to UNESCO (Moseley, 2010). In this paper we start by pondering on the impact that the very development of language science and technology at large is having on language diversity. We will then proceed with this analysis by narrowing the focus of our reflection into a case study in the realm of language resources. With the present paper, we aim at fostering the debate and action on how language scientists and the research activities on language science and technology, including on the development of language resources, can be much more mindful of the language diversity issue and must bring it to the center stage of its mission. It is not only natural and human heritage but also their very own object of research that is being eroded as they are investigating it and because they are investigating it they way that investigation is being undertaken.

2. Monolingual vortex

In the prevailing model for the promotion and funding of research, science progress is mostly driven by the societal priorities identified in the different countries and entities supporting its development. The growth of scientific knowledge is thus asymmetric among different areas, different disciplines inside a given area, different topics inside a given discipline, etc.

Human language science and technology is no exception. As in each country research tends to be prioritized mostly towards its official or predominant language(s), a most substantial asymmetry here is the different research effort devoted to different languages. As a matter of fact, this asymmetry could not be more extreme given that it tells apart one language, viz. English, from all other languages. This was neatly captured in the study of (Mariani and Frankopoulo, 2012), whose findings are summarized in Figure 1, where the research effort devoted to English is five times larger than the effort devoted to the second most researched language, confirmed by an independent study based on a different sample (Rehm and Uszkoreit, 2013), p. 10.

English is the predominant language of the United States, the country that is the world superpower, and of a number of other economically and scientifically highly developed countries, including the United Kingdom, Canada and Australia. All in all, it is the language of around 1/4 of the world GDP — with the second language with the largest share, Chinese, with only half of that GDP value, followed by a long tail of other languages related to economies with GDPs all below half of the Chinese score. But this extreme asymmetry is not explained only by the funding model of science. It results also from how science is produced and to a Matthew effect of accumulated advantage this induces.

At some point or other of their doctoral research on language technology, most students from a non English speaking country had to face the decision on whether they pick their mother tongue or English as the object language of their study. Although English is not their native language, it has many more language resources, processing tools and applications available that can be reused, and many more published research results on which further progress can be built, and this offers a very clear picture: Adopting English as the object language for their research substantially enhances their chances that more new results are produced in less time, and hence their chances of eventually getting more papers accepted for publication, a more visible dissertation and a better career.

We all know many colleagues and students that when faced with this individual choice understandably opted for their immediate best interest and chose English. Regrettably, in collective terms this represents the diverting of the very few resources available in non English speaking countries to support, again, research on English, thus further widening the gulf between advancing the research on English and on the other languages, in detriment of the latter.

This is a self-reinforcing draining effect that is active at many levels that further reinforce each other: When senior researchers decide which research themes to pursue that could be more rewarding for their promotion; when the heads of research units decide which research lines to support that enhances the chances of a better assessment outcome and more future funding for their units, etc.

And all this is strengthened by the fast science funneling effect, where replication and reproduction of previous results are not being produced, accepted or published in almost all venues, not even for English.

It is as if one would have decided, in a counter-factual world, to establish the scientific realm of Biology by researching only one species, or a half dozen of them at best. Certainly these researchers would come across many aspects that we would know that are common and universal to all species and living beings, but they would have no means to figure that out, or even to hypothesize that that was the case given they would have no access to the other thousands of species.

3. Monolingual cyberworld

Along human history, natural language was gone through technological shocks, among which some of the most well known, for instance, are the advent of writing or of the printing press. New technologies have permitted an enhanced usage of languages, allowing to break temporal, spatial and social limitations of face to face communication.

But this usually comes at the cost of a reduction in language diversity. Languages that due to historical or economical circumstances were not technologically prepared or did not receive the benefits of the new technology tended to be abandoned by their speakers. A well known example is the decline and eventual extinction of many languages and dialects that were not used as vehicular languages by the newspapers in previous centuries (Wright, 2016).

Leaving aside the violent cases where languages are banned or their speakers are decimated, languages get extinct because their speakers abandon it in favor of another language. And the causes are basically the same as the ones that lead emigrants to abandon their rural villages to move to big metropolis: The new language is felt to grant more competitive advantages than the one that is getting abandoned, and eventually extinct, in their search for a better life.

Natural language is undergoing a technological shock with unprecedented historical and civilizational consequences. Language technology will allow to overcome further limitations of communication. It will allow speakers that do not share a common language to instantly communicate with each other, supported by automatic translation services. And it will permit to communicate seamlessly with all sort of devices, digital services and artificial agents in natural language.

Even more than in previous technological shocks, language diversity is under the risk of being drastically depleted. The digital world is bringing new disruptive forms of living and working with wide new competitive advantages that cannot be ignored. Natural languages that will be technological prepared will be major channels and instruments to get access to those benefits and will constitute a most relevant competitive advantage. They will be an irresistible attracting pole for speakers of other languages that will not undergo sufficient technological preparation, in yet another instantiation of the Matthew effect of accumulated advantage.

It turns out that language science and technology is neglecting a vast portion of its research object, by neglecting the research on and the technological preparation of the vast majority of the languages. At this point, it is also evident that this risks to contribute for the depletion of its very own research object. The more we funnel our research into one language, or a handful of languages, the more we are contributing for the risk of others to get extinct in the digital age, and thus for further funneling, and eventually locking, ourselves into doing research in one language.

Under the current historical circumstances, ignoring language diversity in our research is not only neglecting a vast portion of our research object and delaying our scientific progress. It is also contributing to reduce language diversity and eventually compromising our chances to get to understand our very research object, human language. Given the way we are doing language science and technology, one could well say that we are contributing to the depletion of our very own research object as we are investigating it. A quite singular situation for the scientific ethos.

4. Multilingual replication

The funneling and depletion effects commented on above can be traced back to the asymmetries underlying how scientific activities happen to be nowadays economically and socially sustained and deployed. These are asymmetries that are common to all scientific areas and disciplines and are inducing all sorts of biases compromising not only the effectiveness but also the integrity of the scientific endeavor.

This discussion has grown in importance as the resources allocated to and societal impact of scientific activities have been expanding (e.g. (Stodden, 2013), (Aarts and others, 2013), to the point that it has crossed the borders of the research world and made its appearance in important mass media and was brought to the attention of the general public (e.g. (Nail, 2011), (Zimmer, 2012), (Begley, 2012), (Begley and Ellis, 2012), (Hiltzik, 2013), (Economist, 2013)). The immediate motivation for this increased interest is to be found in a number of factors, including the realization that for some published results, their replication is not being obtained (e.g. Prinz et al. (2011), Begley and Ellis (2012)); that there may be problems with the commonly accepted reviewing procedures, even besides their possible lack of quality, where deliberately falsified submissions, with fabricated errors and fake authors, get accepted even in respectable journals (e.g. Bohannon (2011)); that the expectation of researchers vis a vis misconduct, as revealed in inquiries to scientists on questionable practices, scores higher than one might expect or would be ready to accept (e.g. Fanelli (2009)); among several others.

Underneath these immediate causes, a number of factors have been pointed out, including 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).

Given the way — together with other colleagues from other disciplines — we are doing science and technology, one could well say that we are contributing to the depletion of the conditions of possibility of our very own research activity. Another quite singular situation for the scientific ethos.

Against this background, it is compelling to advocate that like in other scientific areas, we very much need to foster practices that enhance reproducibility and replicability of research results and bring them to the center stage of our scientific activities, amplifying pioneering initiatives like the 4REAL workshop (Branco et al., 2016). Following the text introducing a new Special Section of the Language Resources and Evaluation journal on reproducibility and replicability (Branco et al., 2017): "Reproduction of results entails arriving at the same overall conclusion(s), as opposed to finding identical values for some measure (Drummond, 2009), (Dalle, 2012), (Buchert and Nussbaum, 2011); that is, to appropriately validate a set of results, scientists should strive to reproduce the same answer to a given research question by different means, possibly by re-implementing an algorithm or evaluating it on a new dataset. Replication has a somewhat more limited aim, typically involving running the exact same system under the same conditions in order to arrive at the same output result."

In a previous occasion, we have motivated this need on the interest of securing the integrity and quality of the research results in our area (Branco, 2013). In the context of the present paper, this need gets further reinforced as a key measure to counteract the funneling and depletion effects that were commented on above and that in the long run appear as self-defeating our own scientific endeavor.

It is important that results obtained when working on some object language(s) are reproduced and replicated with those same language(s) and also with other languages. In the long-term interest of our research area and research subject, it is important that this becomes accepted and encouraged as a first-class citizen practice of our scientific activities.

5. Multilingual diversity

The funneling and depletion effects commented on above can be traced back to overall asymmetries that are common to all scientific areas and disciplines, including ours. This calls for our community to be aligned with and be a major contributor for global correctives initiatives, like paying due attention to replication and reproduction of results in scientific research.

But there are biasing effects that emerge as specific of our area given its particular characteristics and the specific nature of its research subject, and call for responses that should be specific. And for a problem to be addressed and corrected, the first basic requirement is that there is sufficient awareness that it exists.

Anecdotal evidence that in language technology,[1] language diversity is obliterated and that this is not being perceived as an issue, can be found on how titles happen to be chosen for papers in international venues. The few publications whose results are obtained working with an object language different from English typically have an explicit mention to that language in the title. The vast majority of the papers, in turn, which takes English as an object language, makes no reference to English in the title, and many times, not even in the body of the articles. It is compelling to envisage this socially accommodated behavior as a manifestation of a collective unconscious assumption — by all authors, from both sorts of papers alike — that English is "the" natural language by default, and the other languages are just a source of additional exotic or picturesque details.

More seriously than the inessential wording of titles, this bias has been endured by researchers who receive reviews for their papers whose object languages do not include English. If one pays attention to their shared stories during coffee breaks in conferences, one come to realize that more often than not they are questioned by anonymous reviewers whether their results also hold for English (but not, say, for Finnish, Farsi, Hindi, Japanese or any other one of 7 000 languages in the world), or even advised that English should be tried for the paper to be considered mature to be submitted for publication.

Anecdotal aspects aside, the bias this is illustrating has a decisive impact on how our research activities are fostered and our results have been pursued. In this extended abstract, we will focus on one particular example, meant to be illustrative.

As a language resource, WorddNet is a most well known and important asset in language technology. As a multiparty open research initiative, it is a most successful one in our area. And as a case of replication, it is a most prominent one, with an ever growing number of WordNets in construction for particular languages. This is why WorNet offers a telling example that diversity is needed and can be promoted.

[1] With honorable exceptions, like the research communities gathering around LREC/ELRA conferences and only a very few others.

There have been a number of initiatives, including EuroWordNet, MultiWordNet, BalkaNet, etc. (Vossen, 1998), (Pianta et al., 2002), (Tufiş et al., 2000), where concepts that are from different WordNets and are semantically equivalent are co-indexed with each other. As semantic equivalence is a transitive relation, it suffices that each concept, in a particular language/WordNet, is indexed with an equivalent concept, in any other language/WordNet. However, in practice concepts from all languages other than English have been connected to concepts of only one other language, namely English. In practice, no sustained studies, development tools or alternative multilingual ensembles that support a different approach have been pursued.

When there are two equivalent concepts that can be lexically expressed in two languages other than English, but that cannot expressed in English, that equivalence has remained unrecorded. This has funneling effects that once again brings superior development effort and competitive advantage to English. The WordNet for this language has the widest translational homomorphism, built though at the cost of a significant share of the resources deployed for and during the construction of the WordNets for the other the languages. And at the cost that translational equivalence is eventually sub-optimally registered among other languages.

More recently, this funneling effect had been further fostered as a side effect of other initiatives, including Open Multilingual WordNet, BabelNet, etc. (Bond and Paik, 2012), (Navigli and Ponzetto, 2012), whose goal is to gather ensembles of WordNets mapped among themselves, rather than just co-indexing their concepts. As these ensembles start being increasingly used and cited in the literature, the existence of the individual WordNets gets obfuscated. Given these ensembles appear as a convenient one-stop reference, even when only one particular WordNet in them is needed, the former are the favored reference. Citations to individual WordNets are thus vanishing, and with them the incentives and the research productivity indicators that researchers and funding entities need in order to support the continued research on other languages/WordNets.

As a first possible step contributing towards mitigating these funneling effects, we have proposed the undertaking of a Pluricentric Global Wordnet (Branco et al., 2018). But our goal here is not to motivate and present this notion. Rather, WordNet is being offered just as one illustrative case — among possibly many existing ones — of asymmetric biases that may be specific to our field, and to each one of our research topics, and that need to be addressed with specific responses that go on a par with an increased attention to reproduction and replication of results.

Such specific responses are needed to secure and enhance diversity in a wide range of dimensions in our research activities. We need more language diversity in every aspect of our procedures in our scientific activity, ranging from how we set up the reviewing of papers in conferences, to how we conceive our research questions and deploy our priorities around every one of our research topics or subareas, and including crucially how we raise and lobby for the funding of our activities.

Certainly, the need for more language diversity echoes, even if at a different level, the overall need for more diversity in our area, including the diversity in terms of methodological approaches, gender, etc., that have also started to be identified at other venues (Nivre, 2017).

6. Final remarks

In this paper, we presented an analysis aimed at bringing to light two processes that are induced by our research activities in language technology and whose combination are one of the major contributions for the depletion of language diversity. One the one hand, our research focus mostly in one language, English. On the other hand, natural language is undergoing a historical technological shock that is reducing the social and economical competitive advantages for the vast majority of individual languages other than English. Each one of these processes is an instance of and is propelling a powerful Mathew effect of accumulated advantage, which get even further aggravated by the compounding effect of their confluence and the exponential magnifying combination with each other.

Besides raising awareness about this circumstance as a major issue questioning our practice as scientists, in this paper, we also propose measures to mitigate and counteract this unwelcome contribution of language technology for the depletion of its own research subject.

A set of measures results from the alignment with an emerging global trend in science that is urging for greatly increasing the replication and reproduction of research results, for the sake of securing the overall credibility of the scientific knowledge and endeavor. Language technology also needs and benefits from aligning as most and as rapidly as possible with this trend. This process needs and should — and provides a major opportunity — to extend replication and reproduction to languages other than English.

Another set of measures are specific to our area and result from bringing to light and counteracting the non assumed viewpoint that English is the natural language by default. We need more language diversity in our scientific procedures. This should trigger an overall collective process of renewing our activities, ranging from how we set up the reviewing of submitted papers in conferences, to how we conceive our research questions and deploy our priorities around every one of our research topics or subareas, and including crucially how we lobby for the external support to our activities.

Acknowledgment

The present research was partly supported by the Infrastructure for the Science and Technology of the Portuguese Language (CLARIN Língua Portuguesa), and by the ANI/3279/2016 grant.

7. References

Aarts, A. et al. (2013). Estimating the reproducibility of psychological science. *Science*, 349.

Begley, G. and Ellis, L. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, 483:531–533.

Begley, S. (2012). In cancer science, many ”discoveries” don’t hold up. March 28, 2012, online edition. Reuters.

Bohannon, J. (2011). Who's afraid of peer review? *Science*, 341:60–65.

Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71.

António Branco, et al., editors. (2016). *Proceedings of the Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language (4REAL), of LREC2016*. ELRA.

Branco, A., Cohen, K. B., Vossen, P., Ide, N., and Calzolari, N. (2017). Replicability and reproducibility of research results for human language technology: Introducing an lre special section. *Language Resources and Evaluation*, 51:1–5.

Branco, A., Branco, R., Saedi, C., and Silva, J. (2018). Browsing and supporting pluricentric global wordnet, or just your wordnet of interest. In *Proceedings of LREC 2018*.

Branco, A. (2013). Reliability and meta-reliability of language resources: Ready to initiate the integrity debate? In Sandra Kuebler, et al., editors, *Proceedings of the 12th Workshop on Treebanks and Linguistic Theories (TLT12)*, pages 27–36. Bulgarian Academy of Sciences.

Buchert, T. and Nussbaum, L. (2011). Leveraging business workflows in distributed systems research for the orchestration of reproducible and scalable experiments. In Anne Etien, editor, *9ème édition de la conférence MAnifestation des JEunes Chercheurs en Sciences et Technologies de l'Information et de la Communication - MajecSTIC 2012 (2012)*.

Dalle, O. (2012). On reproducibility and traceability of simulations. In *Proceedings of the 2012 Winter Simulation Conference (WSC)*, pages 1–12. IEE.

Drummond, C. (2009). Replicability is not reproducibility: nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*.

Economist. (2013). 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.

Fanelli, D. (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, B. (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, M. (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.

Mariani, J. and Frankopoulo, G. (2012). Language matrices and the language resource impact factor. In *PAROLE Workshop, Lisbon, 18-19 October 2012*.

Christopher Moseley, editor. (2010). *Atlas of the World's Languages in Danger*. UNESCO Publishing, third edition.

Nail, G. (2011). Scientists' elusive goal: Reproducing study results. *The Wall Street Journal*, December 2, 2011, online edition. http://online.wsj.com/news/articles/SB10001424052970 20376480457705 9841672541590.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Nivre, J. (2017). Presidential address ACL 2017: Challenges for ACL. *The 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. https://pt.slideshare.net/aclanthology/joakim-nivre-2017-presidential-address-acl-2017-challenges-for-acl.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). Multiwordnet: Developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*.

Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10:712.

Georg Rehm et al., editors. (2013). *META-NET Strategic Research Agenda for Multilingual Europe 2020*. Springer.

Stodden, V. (2013). Resolving irreproducibility in empirical and computational research. *Institute of Mathematical Statistics Bulletin Online*. http://bulletin.imstat.org/2013/11/resolving-irreproducibility-in-empirical-and-computational-research.

Tufiş, D., Cristea, D., and Stamou, S. (2000). Balkanet: Aims, methods, results and perspectives - a general overview. *Romanian Journal of Information Science and Technology Special Issue*, 7:9–42.

Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer.

Wright, S. (2016). *Language Policy and Language Planning: From Nationalism to Globalisation*. Palgrave Macmillan UK, second edition.

Zimmer, C. (2012). A sharp rise in retractions prompts calls for reform. *The New York Times*, April 16, 2012, online edition. http://www.nytimes.com/2012/04/17/science/rise-in-scientific-journal-retractions-prompts-calls-for-reform.html.

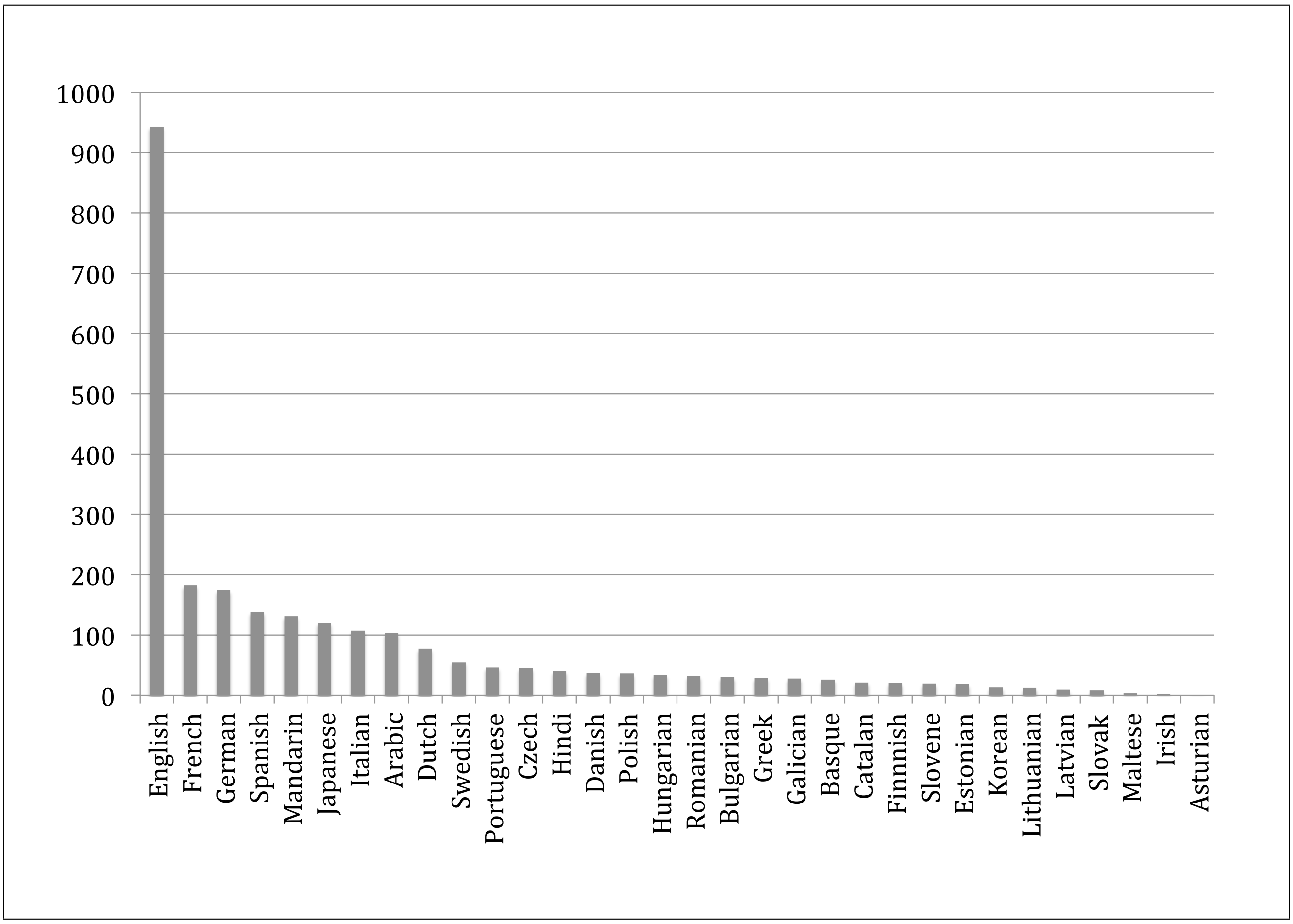

Figure 1: Number of references to data sets and processing tools per language at top scientific conferences between 2010 and in 2012, from (Mariani and Frankopoulo, 2012).