Mind the Gap: Language Data, Their Producers, and the Scientific Process

Tobias Weber
Ludwig-Maximilians-Universität München, Germany

Abstract

This paper discusses the role of low-resource languages in NLP through the lens of different stakeholders. It argues that the current "consumerist approach" to language data reinforces a vicious circle which increases the technological exclusion of minority communities. Researchers' decisions directly affect these processes to the detriment of minorities and practitioners engaging in language work in these communities. In line with the conference topic, the paper concludes with strategies and prerequisites for creating a positive feedback loop in our research benefiting language work within the next decade.

2012 ACM Subject Classification Computing methodologies → Language resources; Computing methodologies → Natural language processing; Social and professional topics → Cultural characteristics; Software and its engineering → Software creation and management; Applied computing → Language translation

Keywords and phrases minority languages, data integration, sociology of technology, documentary linguistics, exclusion

Digital Object Identifier 10.4230/OASIcs.LDK.2021.6

Category Crazy New Idea

1 Introduction

This paper was inspired by the conference organisers' call for "challenging research ideas that [...] you would like to see in ten years from now". One of the major challenges associated to the conference topics is the relationship between different agents: individual speakers providing data, researchers annotating data, computational linguists and data scientists building applications on these data, and the public, i.e., speaking communities, using these applications. What seems like a mono-directional and circular relationship on the macro-level becomes a complex network of interactions on the micro-level. The discussion here shall be led with an interdisciplinary view to present these interactions and the role of three stakeholders: the speakers, the linguists, and the computer scientists. These three groups are to be seen as functional roles which are not mutually exclusive – there can be linguists and computer scientists who also qualify as speakers, or even individuals fulfilling all three roles simultaneously. The paper is written from the intermediary position of the linguist in this list, as it contains my personal experiences as a philologist and curator of language data. While my perspective may not be representative of all linguists, there is a consensus on professional standards in data collection and preparation. Especially for minority languages and recently documented languoids [11], researchers aim to generate sustainable and interoperable data sets for a variety of subsequent uses. As this topic is central to the scientific discourse in documentary linguistics, there is a bulk of literature addressing the issues in language data use. Although the focus of this paper shall be on future developments, it appears imperative to point out that already twenty years ago, before Big Data became a standard paradigm in computer science, linguists highlighted potential conflicts between this approach and the reality faced by those documenting endangered languages. As a consequence, some of these issues have not been fully solved or have worsened over the last decade, making their

3rd Conference on Language, Data and Knowledge (LDK 2021).
Editors: Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch; Article No. 6; pp. 6:1–6:9
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

solution a priority for the next decade. It is important to emphasise that the goal is not simply to force Western paradigms and conceptualisations onto endangered languages, as linguists and computer scientists have warned [8, 22]. The main principles must be the acknowledgement of community agency, reciprocity, and social awareness for the ways our work affects communities. We should not always seek the easy route, where we find large amounts of standardised data, but pay attention to calls for support with data collection and curation.

2 Standardisation – a necessary evil?

In their training, most linguists will learn about technical requirements for data they collect and aim to analyse in their work. This may include file types, encodings, methods of annotation, or tools for the creation of data sets. In many cases, there are accepted standards or conventions for the discipline of linguistics which often follow recommendations from computer scientists. And while these standards help to improve interoperability, durability, and usability in most instances, it would be a fallacy to regard standardisation as a solution in every scenario. Sometimes researchers inherit data sets or projects from colleagues, want to adhere to traditions in their discipline, or must respect requests by a stakeholder. Requiring them to change standards and conventions may create challenges which can only be resolved by investing time and money. In any case, linguists need to be aware of the necessary standards and tools for ensuring compliance with these standards.

I agree with one reviewer that we need to be careful about seeing technology as the central solution in language documentation, where the preoccupation with technological features and facts leads us away from the central role of human relationships and community involvement [14]. This framing commodifies language and data, and has to be seen critical [27]. Instead, we need to take the community and its members into consideration, e.g., under the six Rs presented by Galla and Goodwill [17]: respect, relationality, relevance, responsibility, reciprocity, and resiliency. Yet, even this focus in contemporary documentation and revitalisation efforts will not remove the undesired notion of the commodity from language and data – it has been introduced through globalisation and the subsequent exposure to Western conceptualisations of language and can be found in several indigenous communities around the globe. We must not presume the Western interpretation [8] but, likewise, cannot deny its influence in some communities.

Major challenges are posed by instances where a conversion to a standard cannot be automated, e.g., if annotation, translation, or transcription layers need to be added or altered. These are crucial for many applications in NLP and their quality directly affects the usability of products sold or gifted to the communities. In these cases, data sets need to be manually curated, if the quality of application shall not be impaired. The recent trend of building applications using machine learning or big data approaches requires large amounts of data which needs to conform to particular standards, in addition to being comprehensive. For well-resourced languages (in terms of language resources, time, money, or skilled labour), these obstacles are overcome with seeming ease. Looking at the other end of the scale, under-resourced languages may suffer in multiple respect, relative to these standards. The accumulation of these issues can, subsequently, lead to exclusion, not least in a technological sense [18]. A multi-million token annotated corpus of "gold standard" is comparably rare considering the often cited figure of 7,000 languages and their numerous varieties. As we benchmark applications built for major languages as the state-of-the-art, we set high targets for under-resourced languages which they may not be able to achieve. As a result, we further add to their exclusion. Although there are a range of approaches which aim to create applications from small or unstandardised data sets [2, 3, 5, 15, 16], these papers – in general – summon the "Zero Resource Scenario" criticised by Bird [8], while simultaneously removing speakers and, in some cases, documenters from the line of research. At the same time, these applications cannot be compared in coverage or reliability to applications for major languages and are not solutions against technological exclusion.

In terms of technological exclusion of a language community, we can approximate a trend for a given point in time or interval by calculating the ratio of increases in expected standards and requirements (i.e., size of corpora, quality of translations, consistency of annotation, variety of genres, balance of language data) to the increase in usability (quality and quantity) of language data for the particular languoid, which may include rate of documentation or addition to corpora; measures of quality, balance, or representativeness of data sets and their layers; and the adoption of standards. While these precise calculation of these measures or definition of requirements and standards depend on the desired quality and type of application, this yields $technological\ exclusion = \frac{increase\ in\ language\ resource\ usability}{increase\ in\ required\ standards}$. If the value is below 1, the community is increasingly technologically excluded; if it is at 1 or above the available data sets are sufficient for creating applications of the desired standard, with higher scores generally correlating to better quality and more diverse usage cases of the applications. If we use variables which measure absolute values (e.g., number of tokens in a corpus), we can even calculate how many words, translations, or annotations each speaker, annotator, or researcher must contribute $\frac{(required\ quantity - current\ state)}{community\ size}$. The point about the accumulative nature of exclusion becomes apparent if we accommodate for different capacities in producing language data (e.g., creating data through publications or social media) or ensuring quality (e.g., digitally available data, use of text processing tools). These factor into the equation on the side of the required standards, increasing the burden on each community member. The dynamic nature in standards and corpus development reinforces the trends in technological exclusion, making it more difficult or resource intensive to break the vicious circle. The availability of data conforming to the required standards creates a bottleneck in the creation of applications [1, 28], especially for communities which are already suffering from exclusion.

One reviewer questioned to what extent marginalised communities would want to use their languages in digital spaces. I cannot speak on the behalf of any community, yet, from my experience with European linguistic minorities, I know of positive responses to applications and tools for minority languages. On the one hand, we must not impose Western arguments about rationality, functionality, or instrumental value on these attitudes and decisions [8], as language use may be tied to other domains and reasons. On the other hand, technology falls together with media and telecommunication and forms a domain for language use which can also be a "marker of recognition in the digital realm" [8, p.3509]. In the study of minority language media, substitution with media in other languages is an important measure [24] which can be transferred to technology. Despite the existence of different factors motivating substitution, the *replacement* and not *enrichment* (along the lines of additive and subtractive bilingualism [21]) of technology use can point at structural issues. Grin [20] created a framework of Capacity, Opportunity, and Desire to capture the factors at play in language vitality [7]. A community's desire to use a language in digital spaces should always be matched by appropriate opportunities to do so. Consequently, technology development needs to be coordinated with communities – a mismatch between the desired uses and available opportunities can start the vicious circle of substitution to the detriment of excluded communities.

This issue exhibits features of the Matthew Effect [23], whereby well-resourced languages receive increasingly more and better applications and technological solutions for NLP tasks, while the rest is cast further adrift – in other words, the digital divide is widening. This brings considerations about our research and development of applications into the political sphere, and forces us to give thought to the social implications of our research activities. The decisions we make directly affect the system and may aggravate the negative feedback loop: some low-resource languages have less support for data collection or curation, requiring more time and effort to have data sets conform to steadily rising requirements. As a result, each speaker or researcher associated with the creation, annotation, or curation of language data for these languages has to bear a higher individual burden. The reaction of researchers and community members to this adverse situation can provide us with insights for further developing this field.

3 Reactions and breaking the vicious circle

There are different frameworks which could be used to discuss the reactions shown by community members but, irrespective of a particular context, we can find that the existing structures [19] lead many users to opting for applications in languages other than their preferred one, whether it is for quality (e.g., accuracy of translations), coverage (e.g., specialised vocabulary), or ease of use and accessibility of the other application. This practice of substitution reinforces the adverse structures and facilitates the expression of the relative status or symbolic power of the majority community [9]. Some may show reactance [10] and support the creation of NLP applications for their languoid but, as discussed above, they do so at a high individual burden. We certainly cannot blame those users who opt for more developed resources and applications which cover their needs better, but we must consider the structures which reproduce and reinforce these inequalities and do not provide opportunities for using minority languages. One aspect already addressed critically in the previous section is standardisation and the benchmarking of the best. While it must be our goal to improve our standards and set high targets, we must be aware of the exclusive nature of these reference points which not all languages can meet. By labelling an application "state of the art", we make it desirable and prestigious. At the same time, the combination of high standards/requirements, size of the community, and the status of a language form influence the set of languoids which can successfully aspire to this prestigious technological resources, while the rest suffers from technological exclusion. The vicious circle continues through our individual actions and reactions to the structures – lowering standards, increasing community sizes and user groups, or enhancing the status of a language form are possible solutions but these, generally, lie beyond the remit of the individual researcher.

Looking at the scientific process involving language data in 2004, Nathan and Austin warn that a "consumerist approach" [25] will not support endangered, minoritised, or low-resource languages. There are two important points in that assessment: first, as the authors discuss, there is a metadata gap between collections of language resources in (documentary) linguistics and their subsequent uses in computational linguistics and NLP applications. This disenfranchises, first and foremost, the speakers and consultants who give us permission to record their language use, their stories, and use it for scientific discovery. Certainly, "giving back in ways that are meaningful and valuable to the communities" [6, pp.49–50] is necessary for acknowledging a reciprocal relationship between researchers and the consultants. At the same time, this does not justify extraction and mechanistic decontextualisation (for a criticism of terms like "mining" or "harvesting" in contexts of research on minority languages see

Davis' 2017 article [13]). This extraction does not just affect the original "producers" of data (i.e., speakers and consultants), but also all researchers, annotators, translators who support the creation of language resources and high-quality data sets. I have myself experienced this extraction of data I curated, which I was surprised to see copied on Wikipedia without proper citation or attribution. While this was remedied without any difficulties, it is indicative of the second aspect of the consumerist rhetoric: the omnipresence of data. In our everyday lives, especially in academia, we face large amounts of data – in some instances, we can freely decide which data sets we want to use in our work. Therein lies the problem of the consumerist approach, as most researchers mirror the community members' behaviour outlined above by preferring higher quality, standardised, easily accessible data sets with a large coverage. At the same time, those academics who show reactance and work on low-resource languages face a struggle against evolving standards and expectations which are benchmarked against well-resourced languages. Do we, as academics, also enter an vicious circle – is there a point where we stop caring for minoritised languages and their communities?

I would like to argue with the same frameworks of agency introduced above that we do possess agency as scientists [33]. Especially in instances where our decisions affect communities outside of academia, it is our responsibility to acknowledge this relationship through our decisions. This is relevant in our work with communities who contribute to our research by providing us with data, which we must respect and honour – not just as the documentary linguist who "collects" data (to use the extraction metaphor) but also as the "consumers" of these data who have to acknowledge not only the speakers and consultants but also their colleagues in linguistics or anthropology who enriched data sets and made them usable. All of these stages of "language work" [22], as part of the scientific process, should aim for an appreciative stance of data and its producers. In terms of data citation, there are positive developments towards this goal, e.g., the Tromsø recommendations [4], although I would contend that we must extend that to past publications [30] with a view to preserving the human traces in our work, as also argued by Nathan & Austin [25] and termed "finding the human in the loop" by Bird [8]. Reaching this stance of acknowledging all contributions to a data set can furthermore help to prevent biases [32], as we can, ideally, track and reconstruct links through time and different layers of data [31, discussed under the notion of "Metadata Inheritance"]. These steps are possible for every researcher working with language data and conducting "language work", and leads to a more reciprocal relationship between communities and different groups of scientists.

4 Outlook

The goal of this paper is to highlight different gaps which are widening, as we have been able to observe for the past decade, and which we should aim to close through mindful decisions in our research. These gaps do not only exist between different language communities as technological exclusion – created through different language status, population size, or differences between requirements and capacities – but also between groups of researchers. If the consumerist approach prevails, colleagues working in linguistics departments will be relegated to producers of language data in competition with each other (if we continue with the economicist notion of academia).

This is not a one-way process where consultants provide linguists with data who produce resources and data sets, from which the computer scientist can pick at will and without bearing responsibility. Those who have accepted this responsibility and support communities and colleagues creating data sets deserve utmost respect (e.g., members of the ACL SIGEL, the ComputEL community, or other special interest groups focusing on minority languages). Yet, to overcome the technological exclusion and support low-resource languages, more scholars need to critically assess how their decision-making and pursuit of ever-increasing targets and standards impacts those suffering from structures of exclusion and extraction. If more researchers subscribed to these goals, academia, in general, would be better equipped to support minoritised language communities by facilitating their participation (text processing, keyboards, dictionaries, spell-checking etc.). The precise goals have to be set by the communities and tailored to their needs, respecting their agency [8]; scholars and their institutions have access to the knowledge, the tools, and the funding to realise these projects. In turn, the technological support to minority languages can initiate a virtuous circle, whereby communities can catch up to the standards for more advanced, intelligent, or high-quality NLP applications, if this is seen as a goal for the community members. Considering ways of closing the gap will halt exclusion based on size and status of languages and may set a signal for the community by increasing the prestige of their language. Yet, for this to happen, we need to assess our goals as scholarly community and also address structural problems that encourage and reward those who take the easy path of where the ready-made data sets are. Not only are these options at times easier, research on major languages still attracts more funding, creates opportunities for fast production and output of publications, and tempts researchers with prizes, awards, and reputation. The goal is not to blame and shame colleagues who go down this route or to discredit their work – but to be aware that languages and data are not commodities that can be consumed or exchanged at will, without having implications on language communities. Handling language data and developing NLP applications always brings ethical considerations about social and political consequences of this work, and neither the community of linguists nor the community of computer scientists can escape their responsibility.

Linguistic justice, equal chances for participation, and the acknowledgement of the producers of language data through “giving back” are strong ethical arguments for the stance outlined above. Yet, we can further emphasise the benefits of devoting time to low-resource language to computer scientists. First, some may embrace the multiple challenges which low-resource languages pose and see these contexts as a chance to test new approaches [28]. Second, increasing the amount and quality of language data sets enables testing different approaches across a variety of data, allowing for testing hypotheses about the applicability and quality of applications. For linguistics, it has been argued that minority languages help in testing assumptions and theories by creating diverse data sets [29] and representing some “exceptional types” [26, p.367]. Third, some researchers have argued that minority languages are more likely to have retained rare linguistic features [34], which may not be stable in other languages [12]. Investigating these features can help linguists working on typology, while, simultaneously, allowing NLP scholars to test their applications and tools on languages outside of the commonly used national languages of Indo-European or Standard Average European typology. As a result, the improved and tested applications can be used to create even more data sets, the reversal of the vicious circle. Students and early career researchers should be made aware of these opportunities, while senior researchers, supervisors, and funding bodies can help with creating a conducive environment for starting a virtuous circle.

The perspective presented in this paper is strongly advocating the case of minoritised communities and low-resource languages, and is unlikely to be adopted by all readers. There are numerous arguments in favour of advancing technological standards and computational methods for major languages, which also support minority communities. But the challenge for the next decade is to foster awareness of ways in which our decisions as researchers serve to reproduce and reinforce inequalities and technological exclusion. With 2022–2032 being declared the Decade of Indigenous Languages by the United Nations, the upcoming years are the best time to embrace the needs of indigenous communities and minorities as our research priorities. The first step in overcoming the consumerist approach and considering reciprocity consists in adopting a critical view on the research we conduct and present. A presentation or research outline which – without reflection – follows the rhetorics of "*I downloaded the gold standard corpus with X million tokens*" will not do justice to under-resourced or minoritised languages. These data sets will generally have been created and curated by speakers and linguists, and do not simply exist like products on a goods shelf. Instead, we should consider ways of making our research applicable and usable for other researchers and practitioners doing language work, thereby making it sustainable, as well as supporting communities who suffer from technological exclusion.

References

1 Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947, Valencia, Spain, 2017. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/E17-1088.

2 Željko Agić, Dirk Hovy, and Anders Søgaard. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272, Beijing, China, 2015. Association for Computational Linguistics. doi:10.3115/v1/P15-2044.

3 Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. Multilingual Projection for Parsing Truly Low-Resource Languages. *Transactions of the Association for Computational Linguistics*, 4:301–312, 2016. doi:10.1162/tacl_a_00100.

4 Helene N. Andreassen, Andrea L. Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Koenraad De Smedt, Bradley McDonnell, and the Research Data Alliance Linguistic Data Interest Group. Tromsø recommendations for citation of research data in linguistics, 2019. doi:10.15497/rda00040.

5 Mikel Artetxe and Holger Schwenk. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. doi:10.1162/tacl_a_00288.

6 Peter K. Austin. Communities, ethics and rights in language documentation. In Peter K. Austin, editor, *Language Documentation and Description*, volume 7, pages 34–54. SOAS, London, 2010.

7 Joseph Lo Bianco and Joy Kreeft Peyton. Vitality of heritage languages in the united states: The role of capacity, opportunity, and desire. *Heritage Language Journal*, 10(3):i–viii, 2013. doi:10.46538/hlj.10.3.1.

8 Steven Bird. Decolonising speech and language technology. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, 2020. International Committee on Computational Linguistics. doi:10.18653/v1/2020.coling-main.313.

9 Pierre Bourdieu. *Language & Symbolic Power.* Polity Press, Malden, 1991.

10 Jack Brehm. *A theory of psychological reactance.* Academic Press, New York, 1966.

11 Michael Cysouw and Jeff Good. Languoid, doculect and glossonym: Formalizing the notion "language". *Language Documentation & Conservation*, 7:331–359, 2013.

12 Michael Cysouw and Jan Wohlgemuth. The other end of universals: theory and typology of rara. In Jan Wohlgemuth and Michael Cysouw, editors, *Rethinking Universals*, pages 1–10. De Gruyter Mouton, 2010. doi:doi:10.1515/9783110220933.1.

13 Jenny L. Davis. Resisting rhetorics of language endangerment: Reclamation through indigenous language survivance. In Wesley Y. Leonard and Haley De Korne, editors, *Language Documentation and Description*, volume 14, pages 37–58. EL Publishing, London, 2017.

14 Lise Dobrin, Peter K. Austin, and David Nathan. Dying to be counted: the commodification of endangered languages in documentary linguistics. In Peter K. Austin, editor, *Language Documentation and Description*, volume 6, pages 37–52. SOAS, London, 2009.

15 Long Duong, Trevor Cohn, Karin Verspoor, Steven Bird, and Paul Cook. What can we get from 1000 tokens? a case study of multilingual POS tagging for resource-poor languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 886–897, Doha, 2014. Association for Computational Linguistics. doi:10.3115/v1/D14-1096.

16 Meng Fang and Trevor Cohn. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 587–593, Vancouver, 2017. Association for Computational Linguistics. doi:10.18653/v1/P17-2093.

17 Candace Kaleimamoowahinekapu Galla and Alanaise Goodwill. Talking story with vital voices: Making knowledge with indigenous language. *Journal of Indigenous Wellbeing*, 2(3):67–75, 2017.

18 Duncan Gallie, Serge Paugam, and Sheila Jacobs. Unemployment, poverty and social isolation: Is there a vicious circle of social exclusion? *European Societies*, 5(1):1–32, 2003. doi:10.1080/1461669032000057668.

19 Anthony Giddens. *The Constitution of Society.* University of California Press, Berkeley, 1984.

20 François Grin. *Language Policy Evaluation and the European Charter for Regional or Minority Languages.* Palgrave Macmillan, Basingstoke, 2003.

21 Rodrigue Landry and Réal Allard. Beyond socially naive bilingual education : the effects of schooling and ethnolinguistic vitality on additive and subtractive bilingualism. *NABE Annual Conference Journal*, pages 1–30, 1993.

22 Wesley Y. Leonard. Producing language reclamation by decolonising 'language'. In Wesley Y. Leonard and Haley De Korne, editors, *Language Documentation and Description*, volume 14, pages 15–36. EL Publishing, London, 2017.

23 Robert K. Merton. The Matthew Effect in science. *Science*, 159(3810):56–63, 1968. doi:10.1126/science.159.3810.56.

24 Tom Moring. Functional completeness in minority language media. In Mike Cormack and Niamh Hourigan, editors, *Minority Language Media. Concepts, Critiques and Case Studies*, pages 17–33. Multilingual Matters, Clevedon and Buffalo and Toronto, 2007.

25 David Nathan and Peter K. Austin. Reconceiving metadata: language documentation through thick and thin. In Peter K. Austin, editor, *Language Documentation and Description*, volume 2, pages 179–188. SOAS, London, 2004.

26 Revere Perkins. The covariation of culture and grammar. In Michael Hammond, Edith Moravcsik, and Jessica Wirth, editors, *Studies in Syntactic Typology*, pages 359–378. Benjamins, Amsterdam and Philadelphia, 1988.

27 John E. Petrovic and Bedrettin Yazan. Language as instrument, resource, and maybe capital, but not commodity. In Bedrettin Yazan John E. Petrovic, editor, *The Commodification of Language*, pages 24–40. Routledge, London, 2021.

28 Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601, 2019. doi:10.1162/coli_a_00357.

29 Jan Rijkhoff. Rara and grammatical theory. In Jan Wohlgemuth and Michael Cysouw, editors, *Rethinking Universals*, pages 223–240. De Gruyter Mouton, 2010. `doi:doi:10.1515/9783110220933.223`.

30 Tobias Weber. Can Computational Meta-Documentary Linguistics Provide for Accountability and Offer an Alternative to "Reproducibility" in Linguistics? In Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski, editors, *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OASIcs)*, pages 26:1–26:8, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. `doi:10.4230/OASIcs.LDK.2019.26`.

31 Tobias Weber. Metadata Inheritance: New Research Paper, New Data, New Metadata? In Andrea Mannocci, editor, *Reframing Research Workshop Accepted Papers*. Zenodo, 2020. `doi:10.5281/zenodo.4155362`.

32 Tobias Weber. A philological perspective on meta-scientific knowledge graphs. In Ladjel Bellatreche, Mária Bieliková, Omar Boussaïd, Barbara Catania, Jérôme Darmont, Elena Demidova, Fabien Duchateau, Mark Hall, Tanja Merčun, Boris Novikov, Christos Papatheodorou, Thomas Risse, Oscar Romero, Lucile Sautot, Guilaine Talens, Robert Wrembel, and Maja Žumer, editors, *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium*, pages 226–233, Cham, 2020. Springer International Publishing. `doi:10.1007/978-3-030-55814-7_19`.

33 Tobias Weber and Mia Klee. Agency in scientific discourse. *Bulletin of the Transilvania University of Braşov Series IV: Philology and Cultural Studies*, 13(1):71–86, 2020. `doi:10.31926/but.pcs.2020.62.13.1.5`.

34 Jan Wohlgemuth. Language endangerment, community size and typological rarity. In Jan Wohlgemuth and Michael Cysouw, editors, *Rethinking Universals*, pages 255–278. De Gruyter Mouton, 2010. `doi:doi:10.1515/9783110220933.255`.