Glottocodes: Identifiers Linking Families, Languages and Dialects to Comprehensive Reference Information

Robert Forkel [a,*] and Harald Hammarström [b]

[a] *Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Germany*
E-mail: robert_forkel@eva.mpg.de

[b] *Department of Linguistics and Philology, Uppsala University, Sweden*
E-mail: harald.hammarstrom@lingfil.uu.se

Editors: Julia Bosque-Gil, University of Zaragoza, Spain; Milan Dojchinovski, Czech Technical University in Prague, Czech Republic; Philipp Cimiano, Bielefeld University, Germany
Solicited reviews: Steven Moran, University of Zurich, Switzerland; Jeff Good, University at Buffalo, USA; Menzo Windhouwer, KNAW - Humanities Cluster, Netherlands; One anonymous reviewer

Abstract. Glottocodes constitute the backbone identification system for the language, dialect and family inventory Glottolog (https://glottolog.org). In this paper, we summarize the motivation and history behind the system of glottocodes and describe the principles and practices of data curation, technical infrastructure and update/version-tracking systematics. Since our understanding of the target domain — the dialects, languages and language families of the entire world — is continually evolving, changes and updates are relatively common. The resulting data is assessed in terms of the FAIR (Findable, Accessible, Interoperable, Reusable) Guiding Principles for scientific data management and stewardship. As such the glottocode-system responds to an important challenge in the realm of Linguistic Linked Data with numerous NLP applications.

Keywords: Linguistics, Linked Data, Language Inventory, Linguistic Standards

1. Introduction1

Glottocodes constitute the backbone identification system for the language, dialect and family inventory Glottolog (https://glottolog.org, currently in edition 4.4, [1]). A glottocode consists of four alphanumeric characters (i.e., lowercase letters or decimal digits) and four decimal digits, for example `abcd1234` or `b10b1234`. Glottocodes are complementary to three-letter ISO 639-3 language identification codes (see https://iso639-3.sil.org/) which, however, concern languages only.

In the current release, there are 25,900 glottocodes (8,533 language-level, 4,571 family-level and 12,796 dialect-level).

2. Motivation and History1

Glottocodes were introduced in 2010 by Glottolog collaborator Sebastian Nordhoff, in response to the following requirements:

[*] Corresponding author. E-mail: robert_forkel@eva.mpg.de.

- An ID specifically designed for machine readability, not confusable with an informal or human-directed identifier
- An ID type oblivious to level of linguistic abstraction (idiolect, sociolect, dialect, language, subfamily, family, etc.)
- An ID system for languages that improves on the ISO 639-3 language identifiers in terms of quality, transparency and anchoring

The 8-character long alphanumeric string was designed not to resemble an abbreviation or an easily remembered mnemonic. This was done specifically in order to counter any temptation to capitalize, modify, inflect or translate it, which users might if the ID-string had had a more human-palatable appearance (such as a three-letter mnemonic, a standardized name or the like).

Glottolog had adopted a doculect-based approach[1] for organizing concrete attestations of languages as recorded by bibliographical references. This means that language data (ultimately emanating from idiolects of specific speakers) recorded in different publications are grouped into successively larger conglomerates such as subdialects, dialects, languages, subfamilies and families [7]. Earlier approaches had always sought to tie the identity of a set of data to a specific level such as language or dialect, such that if the set of data did not change but the level changed (e.g., a language reconsidered a dialect or vice versa) the ID had to change as well (e.g., [8]). Given controversies over language and dialect status, along with our incomplete understanding of the language situation for many minority languages around the world, such level changes are actually not uncommon. With the level-independent ID-system of glottocodes, merely a level-attribute, but not the ID itself would have to change in such cases. The level-neutral term for the denotation of a glottocode is *languoid* [2].

Finally, a decade or longer ago, the quality and transparency of the ISO 639-3 standard for languages was problematic and an alternative was clearly needed [9]. ISO 639-3[2] aims at complete coverage of all natural languages and thus supersedes the earlier ISO 639-2 and ISO 639-1 standards. SIL International is the registration authority for ISO 639-3 and also publish the Ethnologue language catalogue [10]. ISO 639-3 carries little metadata and/or justification for its entries as part of the standard but the information in the corresponding entries in Ethnologue is in practice often taken to substitute as such. Fortunately, the quality and transparency has improved in the last decade so that the discrepancy between language-level glottocodes and ISO 639-3 codes is diminishing (but not completely eradicated, see Section 3.3).

Glottolog was initiated by Harald Hammarström, Sebastian Nordhoff and Martin Haspelmath and is now run by a group of editors[3]. Editors are chosen per release based on unanimous approval of the previous group of editors.

3. Glottolog data curation1

As explained above, glottocodes are the identifiers for *languoids* — the main objects of the data curated in Glottolog. Glottolog also curates bibliographical references ('langdoc') in much the same way, though this is not the focus of the present paper.

In the following, we briefly describe the infrastructure and framework pioneered for the Glottolog data for data curation and publication ([11] and [12]).

- All data is stored in UTF-8 encoded text files (with consistency ensured by the `pyglottolog` software package ([13])).
- Thus, collaboration and curation workflows can make use of `git`, a distributed version control system to track history of changes and provenance.
- The master copy of the Glottolog `git` repository is hosted with GitHub at https://github.com/glottolog/glottolog for curation (but due to the nature of git repositories — where all essential metadata is part of every copy of the repository — this does not put the data at the mercy of GitHub).

[1] The term 'doculect' emanates from [2] but the general approach has been used by many earlier authors, notably [3–6].

[2] See https://iso639-3.sil.org/.

[3] https://github.com/glottolog/glottolog/blob/master/CONTRIBUTORS.md

– Released versions of the repository are published and archived with Zenodo (https://zenodo.org).

This setup does not only provide a stable management system for all information shared between the Glottolog editors, but also a collaborative environment which allows involving the wider community. Somewhat similar to — but less formal than — ISO 639-3 change requests, Glottolog users can make use of GitHub issues to indicate errors or request inclusion of new languoids[4], or even submit pull requests with data corrections[5].

Glottolog aims to share this data in an open and FAIR ([14]) way. Stepping through the FAIR Guiding Principles for scientific data management and stewardship[6] will shed light on the details.

3.1. Glottolog data is findable2

Glottolog is a well established language catalog as evidenced by more than 600 citations of editions of Glottolog such as "Glottolog 4.0" in the scholarly literature (according to Google Scholar). The Glottolog data repository lists 20 contributors in addition to the Glottolog editors (and not including users opening issues, see https://github.com/glottolog/glottolog/graphs/contributors) — pointing to a healthy, collaborative user community.

> *Glottolog data are registered and indexed in a searchable resource.*

Glottolog data is also well indexed in relevant catalogues: The first point of contact for many users is the Glottolog web application at https://glottolog.org — not at least because it is well indexed by Google and other search engines. But Glottolog data is also harvested by OLAC (Open Language Archives Community) and is listed in the OLAC catalogue as the archive with the highest number of distinct languages (see http://www.language-archives.org/metrics/glottolog.org). Finally, Glottolog data releases can be found on Zenodo, and consequently wherever Zenodo metadata is indexed.

> *Glottolog data are assigned a globally unique and persistent identifier.*

All languoids in Glottolog are unambiguously identified via glottocodes. These glottocodes are transparently associated with URLs in the glottolog.org domain, turning them into globally unique identifiers. Each release of Glottolog is identified by the DOI assigned by Zenodo.

> *Metadata clearly and explicitly include the identifier of the data they describe.*

CLDF — one of the dissemination formats of Glottolog — is designed to allow for explicit linking of metadata to identifiers. The underlying mechanism to do this is described in [15], and the sematics are provided through the CLDF Ontology (see https://cldf.clld.org/v1.0/terms.rdf).

3.2. Glottolog data is accessible2

> *Glottolog data are retrievable by their identifier using a standardised communications protocol.*

Zenodo (and the metadata associated by Zenodo with the DOI assigned to data releases) guarantees that data is retrievable using the standard protocol associated with DOIs.

For each languoid, the CLDF/CSVW data associates an HTTP ([16]) URL, which is resolvable via the Glottolog web application.

3.3. Glottolog data is interoperable2

> *Glottolog data use a formal, accessible, shared, and broadly applicable language for knowledge representation.*

Glottolog aims at integration with the Semantic Web at large and the Linguistic Linked Data initiative in particular.

At the most fundamental level this means resource URLs — aka URLs for Glottocodes. These resource URLs are not only usable as universally unique identifiers, but are also resolvable through the Glottolog web application. HTTP status codes ([16]) returned by the web application signal the status of Glottocodes as follows:

200 OK for active codes
410 GONE or `301 MOVED PERMANENTLY` for retired Glottocodes
404 NOT FOUND for invalid codes

[4] E.g. https://github.com/glottolog/glottolog/issues/646
[5] E.g. https://github.com/glottolog/glottolog/pull/648
[6] https://www.go-fair.org/fair-principles/

The web application also provides several serializations of RDF ([17]) representations of languoid data. These serializations can be retrieved using standard content negotiation mechanisms such as using `ACCEPT` HTTP headers.

While these efforts provide convenient integration with the "living" Semantic Web, Glottolog also aims at interoperability for its archived, longterm available datasets. To this end, Glottolog data is serialized as a CLDF Structure Dataset ([18]). The CLDF standard ([19], [20]) does not only provide interoperability with other CLDF datasets, but — due to being built on the W3C's "CSV on the Web" recommendation ([15] and [21]) — also allows automatic conversion to RDF ([22]).

Glottolog data use vocabularies that follow FAIR principles.

CLDF bundles data with structured, machine readable, semantic web-ready metadata. Since CLDF metadata is encoded in JSON-LD ([23]), the data can be marked up using standard ontologies such as Dublin Core, DCAT (https://www.w3.org/ns/dcat#) and PROV (https://www.w3.org/ns/prov#).

Glottolog data include qualified references to other (meta)data.

Thanks to improvements in the curation of ISO 639-3 language identifiers during the last decade, ISO 639-3 codes and language-level glottocodes are one-to-one interchangeable for the vast majority of cases, and the differences are few enough that a specific comment explaining the differences are given in each of the remaining cases on Glottolog. In fact, Glottolog aims at covering all valid ISO 639-3 codes to provide a full mapping, but typically there is a time lag of a couple of months between additions to ISO 639-3 and a Glottolog release addressing these changes. There remains a principled difference in anchoring in that the denotation of a glottocode in Glottolog is defined by the data and information in the references tied to it. The references are associated in Glottolog in such a way that the referenced data and information is enough to distinguish the languoid from all other languoids. Strictly speaking, the ISO 639-3 standard provide no definition or justification of the recorded entries. In Ethnologue [10] — the reference for most of the ISO 639-3 codes — each entry has metadata such as geographical information, name(s), speaker numbers and classification which presumably defines the language, but no actual or referenced data from the denoted language. Unfortunately, it is not so that metadata information is in all cases enough to identify its denotation. Language names are notoriously ambiguous and the case of language-shifting ethnic groups is particularly tricky, as most metadata (speaker numbers, geography, name) is not sufficient to disambiguate between the original and substituted language.

Glottolog, and in particular individual languoids, are also well-linked from Wikipedia. In particular, practically all language- and family-level languoids are referenced in Wikipedia. These Wikipedia links translate to Wikidata links (e.g. https://www.wikidata.org/wiki/Q31746) which in turn provide links to other language identification schemes such as ISO 639-2 (which are arguably less important than ISO 639-3 in the contexts where Glottolog is most used).

3.4. Glottolog data is reusable2

Glottolog data are released with a clear and accessible data usage license.

Glottolog data is release under a CreativeCommons CC-BY-4.0 license.

Glottolog data are associated with detailed provenance.

Like most large-scale databases, parts of Glottolog data are aggregated from various sources. Glottolog tries to be transparent about this, e.g. by

- providing references for all classification proposals[7]
- providing references for all endangerment assessments[8]
- describing the provenance of the bibliography[9]

Glottolog data meet domain-relevant community standards.

[7] E.g. https://github.com/glottolog/glottolog/blob/v4.4/languoids/tree/indo1319/anat1257/luvi1234/cari1274/md.ini#L59-L63

[8] See https://github.com/glottolog/glottolog/blob/v4.4/config/aes_sources.ini

[9] See https://github.com/glottolog/glottolog/blob/v4.4/references/BIBFILES.ini

We already described the relation between glottocodes and ISO 639-3 language codes. Arguably, the transparent mapping between the two, which Glottolog provides, is the most important contribution towards meeting domain-relevant standards.

But as explained above, Glottolog also

- caters to the LLD community, by meeting Semantic Web standards, e.g. re-using ontologies like GOLD[10] to identify languoid levels and Lexvo.org[11] to identify ISO 639-3 codes in Glottolog's RDF formats,
- serves the OLAC community, by implementing the OAI-PMH data provider specification ([24]), thereby allowing harvesting through OLAC,
- helps researchers in descriptive and comparative inguistics to inform their analyses using Glottolog metadata, by making this data accessible as CLDF dataset,
- provides the NLP community with the means necessary to follow the "Bender Rule" [25] of always identifying the language(s) (or language varieties) involved in NLP research

4. Policies governing glottocode assignment1

Glottolog aims to be complete with respect to all assertable L1 languages[12] in the real world, so all languages in the world (as far as this is understood at the time of a certain release) have a language-level glottocode. Glottolog makes a classification decision for all language-level languoids so the family-level inventory is complete in the sense of exhausting the languages of a given release.

Glottolog also classifies dialects insofar as it attaches them to exactly one language-level languoid. But the inventory of dialects (varieties of a language), non-L1 languages (artificial languages, speech registers, pidgins) and non-assertable languages (putative languages for which there is insufficient data to decide if they are different from all other languages) and putative families (hypotheses about family relationships that have appeared in the literature) is growing but still far from complete. The world may contain more of any or all of these entities without a necessary reflection in a glottocode. Genuine completeness with respect to these categories is deemed practically (if not theoretically) impossible.

For these reasons, Glottolog accounts for any changes to the language-level inventory between two releases, i.e. language-level glottocodes of the previous release will always be valid glottocodes in the next. So if something was deemed a real-world language, a user can follow any changes to that assertion. If a language-level languoid was completely erroneous, it is moved to the Bookkeeping category. If it is promoted/demoted to a family/dialect, it retains its glottocode but changes its level accordingly, but from that point on it ceases to be "protected" by its language-level status, so may be retired in the next release[13]. In contrast, the family-level and dialect-level glottocodes are not "protected" and may be removed from the inventory between releases. Since they do not necessarily reflect a real-world entity like an L1 language, it cannot systematically be explained what "happened" to them, e.g., if they never really existed. However, some tracking possibilities are always guaranteed because both families and dialects are linked to language-level languoids. If a family-level glottocode disappears, it is possible to check which language-level languoids it covered and to check which family/ies they are now associated with and if a dialect-level glottocode disappears, it is possible to check which language-level languoid it pertained to and to check which dialects are now associated with it. Furthemore, the git-versioning and the structure of the index allows a quick location of the specific pull-request associated with the removal/appearance of a glottocode (see Section 5).

Glottocodes are not recycled — for new entities, completely new glottocodes are assigned (retired codes are not re-used/re-purposed). Hence, all glottocodes that have ever appeared are either active or retired.

[10] http://purl.org/linguistics/gold/

[11] http://lexvo.org/

[12] See https://glottolog.org/glottolog/glottologinformation for an explanation of these criteria.

[13] This process works similar to the way deprecation (https://en.wikipedia.org/wiki/Deprecation) is used in software development to provide limited backwards compatibility.

5. Glottolog versioning1

We already pointed out that Glottolog data is versioned and released periodically (aiming at a bi-annual release frequency). Each such Glottolog release is self-contained, i.e. does not reference any base data, but instead includes it. Thus, when linking other resources to Glottolog, one should always specify the particular target version.

Glottolog follows a semantic-versioning scheme[14] for data:

– Resources using Glottolog should always target the highest patch version of a particular minor version. This should not break any processing code, but may correct errata.
– Upgrading resources to a new minor version may change data/links, but should not break processing code.
– Upgrading to a new major version may break processing code, i.e. the data structure may change.

The Glottolog version history can be explored in two ways: The Glottolog web application resolves resource URLs of obsolete languoids as follows:

```
$ curl -I https://glottolog.org/resource/languoid/id/awun1244
HTTP/1.1 301 Moved Permanently
Date: Fri, 08 Jan 2021 12:06:32 GMT
Content-Type: text/html; charset=UTF-8
Location: https://glottolog.org/files/glottolog-4.0/awun1244.html
```

Where the HTML page at https://glottolog.org/files/glottolog-4.0/awun1244.html provides context about the languoid and versions when it was still active. In the case of obsolete dialect-level languoids — such as Awuna — this typically allows determining the parent language-level languoid, which will always be part of the current release.

Alternatively, since Glottolog data is curated as `git` repository, we can use the `git` software to inspect the history. In the case of Awuna (`awun1244`), e.g., we learn that it was removed during a "cleanup" of the classification of the Gbe sub-family:

```
 $ git log --all --full-history -- "**/awun1244/md.ini"
commit 923cd9eb21f27bc9ae0797a315dd48fd009e53c4
Author: d97hah <harald@bombo.se>
Date:   Thu Jul 25 14:50:11 2019 +0200
    Gbe (#387)
    * Gbe resolved clf + move of some Mixed languages
    * fix to phoenician-punic+ugaritic
    * New hh.bib + clf ref updates
    * synched refs
```

[14] https://semver.org/

6. Conclusion1

We have described the practices and principles for glottocodes as the identificational system for the languages, dialects and families of the world including data curation, technical infrastructure and update/version-tracking systematics. The resulting data observes the crucial aspects of the FAIR (Findable, Accessible, Interoperable, Reusable) Guiding Principles for scientific data management and stewardship. As such the glottocode-system responds to an important challenge in the realm of Linguistic Linked Data with numerous NLP applications.

References

[1] H. Hammarström, R. Forkel, M. Haspelmath and S. Bank, glottolog/glottolog: Glottolog database 4.4, Zenodo, 2021. doi:10.5281/zenodo.4761960.

[2] M. Cysouw and J. Good, Languoid, Doculect, Glossonym: Formalizing the notion "language", *Language Documentation and Conservation* **7** (2013), 331–359.

[3] W. Schmidt, Gliederung der australischen Sprachen, *Anthropos* **7, 7, 7, 8, 9, 12/13, 12/13** (1912, 1912, 1912, 1913, 1914, 1917/1918, 1917/1918), 230–251, 463-497, 1014-1048, 526-554, 980-1018, 437-493, 747-817.

[4] Č. Loukotka, *Classification of the South American Indian Languages*, Reference Series, Vol. 7, Los Angeles: Latin American Center, University of California, 1968.

[5] G. van Bulck, *Les recherches linguistiques au Congo Belge: résultats acquis, nouvelles enquêtes à entreprende*, Mémoires de l'Institut Royal Colonial Belge, Bruxelles, Vol. 16, Bruxelles: Institut Royal Colonial Belge, Bruxelles, 1948.

[6] J. Good and C. Hendryx-Parker, Modeling Contested Categorization in Linguistic Databases, in: *Proceedings of the EMELD 2006 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art. Lansing, Michigan. June 20-22, 2006*, Lansing, Michigan: E-MELD, 2006, pp. 1–22.

[7] S. Nordhoff and H. Hammarström, Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources, in: *Proceedings of the First International Workshop on Linked Science 2011*, T. Kauppinen, L.C. Pouchard and C. Keßler, eds, CEUR Workshop Proceedings, Vol. 783, CEUR, 2011, pp. 1–7. http://iswc2011.semanticweb.org/fileadmin/iswc/Papers/Workshops/LISC/nordhoff.pdf.

[8] M. Mann and D. Dalby, *A thesaurus of African languages: a classified and annotated inventory of the spoken languages of Africa with an appendix on their written representation*, London: Hans Zell, London, 1987. ISBN 9780905450247.

[9] H. Hammarström, Ethnologue 16/17/18th editions: A comprehensive review, *Language* **91**(3) (2015), 723–737, Plus 188pp online appendix..

[10] D.M. Eberhard, G.F. Simons and C.D. Fennig, *Ethnologue: Languages of the World*, 24 edn, Dallas: SIL International, 2021. http://www.ethnologue.com.

[11] R. Forkel, Glottolog 3.0 released, 2017. https://clld.org/2017/03/29/glottolog-3-0.html.

[12] R. Forkel, Glottolog 3.0 – A collaborative, versioned catalog of languages and dialects, 2016. https://clld.org/docs/poznan/glottolog-3-0.pdf.

[13] R. Forkel, S.J. Greenhill and C. Rzymski, glottolog/pyglottolog: Glottolog API, Zenodo, 2020. doi:10.5281/zenodo.2620249.

[14] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* **3**(160018) (2016), 1–9.

[15] J. Tennison, G. Kellogg and I. Herman, Model for Tabular Data and Metadata on the Web, Technical Report, World Wide Web Consortium (W3C), 2015. http://www.w3.org/TR/tabular-data-model/.

[16] R. Fielding and J. Reschke, Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content, RFC, 7231, RFC Editor, 2014. ISSN 2070-1721. https://www.rfc-editor.org/rfc/rfc7231.txt.

[17] F. Manola and E. Miller, RDF Primer, Technical Report, W3C, 2004. https://www.w3.org/TR/2004/REC-rdf-primer-20040210/.

[18] H. Hammarström, R. Forkel, M. Haspelmath and S. Bank, glottolog/glottolog-cldf: Glottolog database 4.4 as CLDF, Zenodo, 2021. doi:10.5281/zenodo.4762034.

[19] R. Forkel, J.-M. List, S.J. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G.A. Kaiping and R.D. Gray, Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics, *Scientific Data* **5**(180205) (2018), 1–10. doi:https://doi.org/10.1038/sdata.2018.205. https://www.nature.com/articles/sdata2018205.

[20] R. Forkel, J.-M. List, M. Cysouw and S.J. Greenhill, CLDF 1.0, Technical Report, Max Planck Institute for the Science of Human History, Jena, 2017. doi:10.5281/zenodo.1117644.

[21] R. Pollock, J. Tennison, G. Kellogg and I. Herman, Metadata Vocabulary for Tabular Data, Technical Report, World Wide Web Consortium (W3C), 2015. https://www.w3.org/TR/tabular-metadata/.

[22] J. Tandy, I. Herman and G. Kellogg, Generating RDF from Tabular Data on the Web, Technical Report, World Wide Web Consortium (W3C), 2015. https://www.w3.org/TR/csv2rdf/.

[23] M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, P.-A. Champin and N. Lindström, JSON-LD 1.1 – A JSON-based Serialization for Linked Data, Technical Report, World Wide Web Consortium (W3C), 2020. https://www.w3.org/TR/json-ld/.

[24] C. Lagoze, H.V. de Sompel, M. Nelson and S. Warner, The Open Archives Initiative Protocol for Metadata Harvesting, Technical Report, Open Archives Initiative, 2002. http://www.openarchives.org/OAI/openarchivesprotocol.html.

[25] E.M. Bender, On achieving and evaluating language independence in NLP, *Linguistic Issues in Language Technology* **6** (2011), 1–26.

[26] D. Hovy and S.L. Spruit, The Social Impact of Natural Language Processing, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 591–598. doi:10.18653/v1/P16-2096. https://aclanthology.org/P16-2096.