



VOLUME 12 ISSUE 2
DECEMBER 2024

Journal of Language Modelling

VOLUME 12 ISSUE 2
DECEMBER 2024

Editorials

Computational approaches to morphological typology 271
Micha Elsner, Sacha Beniamine

Articles

Alignment everywhere all at once:
Applying the late aggregation principle
to a typological database of argument marking 287
*David Inman, Alena Witzlack-Makarevich,
Natalia Chousou-Polydouri, Melvin Steiger*

Zero marking in inflection: A token-based approach 349
Laura Becker

An analogical approach to the typology of inflectional complexity 415
Matías Guzmán Naranjo

Corpus-based measures
discriminate inflection and derivation cross-linguistically 477
Coleman Haley, Edoardo M. Ponti, Sharon Goldwater



JOURNAL OF
LANGUAGE MODELLING

ISSN 2299-8470 (electronic version)

ISSN 2299-856X (printed version)

<http://jlm.ipipan.waw.pl/>

MANAGING EDITOR

Adam Przepiórkowski IPI PAN

GUEST EDITORS OF THIS SPECIAL ISSUE

Micha Elsner The Ohio State University

Sacha Benjamine University of Surrey

SECTION EDITORS

Elżbieta Hajnicz IPI PAN

Małgorzata Marciniak IPI PAN

Agnieszka Mykowiecka IPI PAN

Marcin Woliński IPI PAN

STATISTICS EDITOR

Łukasz Dębowski IPI PAN



Published by IPI PAN


Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland

Circulation: 50 + print on demand

Layout designed by Adam Twardoch.

Typeset in X₃L^AT_EX using the typefaces: *Playfair*
by Claus Eggers Sørensen, *Charis SIL* by SIL International,
JLM monogram by Łukasz Dziedzic.

*All content is licensed under
the Creative Commons Attribution 4.0 International License.*

 <http://creativecommons.org/licenses/by/4.0/>

EDITORIAL BOARD

Steven Abney University of Michigan, USA

Ash Asudeh University of Rochester, USA

Igor Boguslavsky Technical University of Madrid, SPAIN

Paul Boersma University of Amsterdam, THE NETHERLANDS

Olivier Bonami Université Paris Cité,
Laboratoire de linguistique formelle, CNRS, FRANCE

Robert D. Borsley Professor Emeritus, University of Essex;
Honorary Professor, Bangor University, UNITED KINGDOM

António Branco University of Lisbon, PORTUGAL

David Chiang University of Notre Dame, USA

Dan Cristea University of Iași, ROMANIA

Berthold Crysmann Université Paris Cité,
Laboratoire de linguistique formelle, CNRS, FRANCE

Jan Daciuk Gdańsk University of Technology, POLAND

Łukasz Dębowski Institute of Computer Science,
Polish Academy of Sciences, POLAND

Mary Dalrymple Professor Emerita, University of Oxford, UNITED KINGDOM

Anette Frank Universität Heidelberg, GERMANY

Claire Gardent LORIA, CNRS and Université de Lorraine, FRANCE

Jonathan Ginzburg Université Paris Cité, Laboratoire de linguistique
formelle, CNRS; Laboratoire d'Excellence LabEx-EFLt, FRANCE

Thomas Graf Stony Brook University, UNITED STATES

Stefan Th. Gries University of California, Santa Barbara, USA;
Justus Liebig University Giessen, GERMANY

Adam Jardine Rutgers Department of Linguistics, UNITED STATES

Heiki-Jaan Kaalep University of Tartu, ESTONIA

Laura Kallmeyer Heinrich-Heine-Universität Düsseldorf, GERMANY

Jong-Bok Kim Kyung Hee University, Seoul, KOREA

Kimmo Koskenniemi Professor Emeritus, University of Helsinki, FINLAND

Jonas Kuhn Universität Stuttgart, GERMANY

Alessandro Lenci University of Pisa, ITALY

John J. Lowe University of Oxford, UNITED KINGDOM

Ján Mačutek Comenius University, Bratislava, SLOVAKIA

Igor Mel'čuk Professor Emeritus, University of Montreal, CANADA

Richard Moot CNRS, LIRMM, University of Montpellier, FRANCE

Glyn Morrill Technical University of Catalonia, Barcelona, SPAIN

Stefan Müller Humboldt Universität zu Berlin, GERMANY

Mark-Jan Nederhof University of St Andrews, UNITED KINGDOM

Petya Osenova Sofia University, BULGARIA

David Pesetsky Massachusetts Institute of Technology, USA

Maciej Piasecki Wrocław University of Science and Technology, POLAND

Christopher Potts Stanford University, USA

Agata Savary University of Paris-Saclay, FRANCE

Sabine Schulte im Walde Universität Stuttgart, GERMANY

Stuart M. Shieber Harvard University, USA

Mark Steedman University of Edinburgh, UNITED KINGDOM

Stan Szpakowicz Professor Emeritus, University of Ottawa, CANADA

Shravan Vasishth Universität Potsdam, GERMANY

Aline Villavicencio Institute for Data Science and Artificial Intelligence
University of Exeter; University of Sheffield, UNITED KINGDOM

Veronika Vincze University of Szeged, HUNGARY

Shuly Wintner University of Haifa, ISRAEL

Zdeněk Žabokrtský Charles University in Prague, CZECH REPUBLIC

Computational approaches to morphological typology

Micha Elsner¹ and Sacha Beniamine²

¹ The Ohio State University

² University of Surrey

INTRODUCTION

1

Theories of morphology pertain to the lexicons of languages: what forms of words exist, how they relate to one another, and what they mean. To refine and test such theories, morphologists require high-quality information about lexicons, and where they posit particular learning mechanisms, these naturally operate on lexical knowledge to make their predictions. The size of a natural language lexicon, with its various quirks and irregularities in form and frequency, lends itself naturally to a databasing approach, and morphologists have a long history of productive engagement with computation.

The classification of languages into morphological types constitutes one of the earliest attempts to linguistic typology (von Schlegel 1818). As soon as 1960, Greenberg sought to objectivise these types by calculating indexes on corpora. In the past two decades, different strands of multi-variate morphological typology have converged to set the scene for scaling up morphological typology. The program of Canonical Typology (see among others Corbett 2005; Brown *et al.* 2012; Corbett 2023) has contributed to map out the space of typological variation in morphology and at its interfaces. Simultaneously, the program of Autotypology (see among others Bickel and Nichols 2002; Bickel *et al.* 2022; Witzlack-Makarevich *et al.* 2022) has supported the creation of large, interconnected typological databases, flexible enough to support diverse typological investigations. In inflection,

the conversation on morphological complexity shifted gradually from the search of natural limits on morphological complexity (such as the Paradigm Economy Principle or the No Blur Principle, see Carstairs 1987; Cameron-Faulkner and Carstairs-McCarthy 2000), to the careful measure of this complexity, accompanied with a general turn towards Word & Paradigm approaches (Stump and Finkel 2013). Relying on quantitative analysis, Ackerman and Malouf (2013) describe two kinds of morphological complexity: enumerative (E-complexity) measuring how ‘large’ the system is and integrative (I-complexity) measuring its inter-predictability. Cotterell *et al.* (2019) conjecture that E- and I-complexity trade off against one another, so that languages with larger paradigms are easier to predict, and finds support for this proposal in a dataset of 36 Unimorph languages.

Two great endeavours underpin computational approaches to morphological typology: the elaboration of computational databases and the modelling of morphological systems based on this data. Constructing a computational database for a single language is a serious undertaking, so early studies often restricted themselves to a single language or a handful of related ones. Typological surveys, on the other hand, might be biased in the regions or language families they were able to cover, or might be forced to rely on unstandardised descriptions of different languages in which underlying similarities might be concealed by choices in analysis. Recent trends in morphological typology are striving to close this gap. Larger databases, representing more languages and phenomena, or connected together through standardisation and linked data, allow researchers to scale their modelling studies beyond the best-studied European languages. At the same time, modelling contributes to the standardization of typological description, by defining replicable measurements of theoretical constructs like ‘zero markers’, ‘number of inflection classes’ or ‘inflection vs. derivation.’ Thus, database construction and modelling are potentially synergistic activities which can feed one another, expanding our coverage of human languages while ensuring that our analytical constructs are valid.

While early morphological projects used small ad-hoc datasets or larger resources covering only one or two languages, recent projects have drawn on larger standardised resources. On the one hand, databases of inflected or derived forms document entire un-analyzed

morphological systems. For example, Batsuren *et al.* (2022) provides structured lexical data for 169 languages in a unified format, and the Paralex standard (Beniamine *et al.* 2023) provides conventions to encode rich linguistic information concerning such inflectional resources. These databases of forms allow researchers to test conjectures about the statistics of lexicons at scale. On the other hand, databases of languages provide coded examples of a single phenomenon across many languages (Haspelmath *et al.* 2013; Bickel and Nichols 2002; Skirgård *et al.* 2023).

THE PAPERS IN THIS ISSUE

2

The first paper of this volume describes a novel cross-linguistic database, following the Autotyp approach. The three subsequent papers follow in the tradition of Ackerman and Malouf (2013) by proposing new models.

*Inman & al: Alignment everywhere all at once:
Applying the late aggregation principle
to a typological database of argument marking*

2.1

Inman *et al.* (2024) introduce the ATLAS Alignment Module, a typological database of argument marking at the morpho-syntactic interface, for languages of North and South America. The database is meant to capture the considerable language-internal variation in argument marking. It focuses on main declarative clauses with verbal predication and positive polarity. To a large extent, it conforms to the principles of Autotyp: it is *modular*, with each module covering a specific typological domain; variables and their values were kept open throughout coding (*autotypology*), ensuring detailed and faithful encoding. It enables *late aggregation*, where generalisations are not primary, but instead derived from data encoded at a granular level. Finally, it relies on *exemplars*. The database documents three argument roles (S, A, P) defined by semantics. Across languages, these roles can align together in various fashion, leading to basic alignment types.

For example, in a nominative-accusative alignment, roles S and A are aligned together, and distinctly from P, whereas in ergative-absolutive alignments, S and P are aligned together and contrast with A. Argument selectors are the devices by which arguments can be treated identically or differently, through either morphological marking or syntactic behaviour. Inman *et al.* (2024) focus on two types of selectors: flagging, which pertains to case marking and adposition within NPs, and indexing, which concerns verbal marking and agreement within clauses. The database is distributed in CLDF format, as a set of csv tables. It documents specific alignment contexts, the selectors involved, as well as the languages documented, the database source, and information aggregated automatically concerning references and alignment.

In short, Inman *et al.* (2024) present a wealth of precise data on alignment which can be aggregated at any documented level. It will enable testing numerous typological hypotheses, definitions, and operationalisations, much beyond those which were considered by the database authors.

2.2

Becker: Zero marking in inflection: A token-based approach

Becker (2024) tackles the challenge of observing the invisible. What is the typological distribution of zero markers? Do they behave like short markers, which, for reasons of coding efficiency, tend to be more frequent and predictable than longer markers (Zipf 2013; Greenberg 1966; Haspelmath 2008)? Becker surveys adjectival, nominal and verbal systems from 114 languages across six macro-areas. The data is derived from Unimorph (Kirov *et al.* 2016, 2018; McCarthy *et al.* 2020; Batsuren *et al.* 2022), with pre-processing to improve data quality and comparability, including conversion of some datasets to phonemic representations. Zero marking is unfortunately difficult to distinguish in a principled manner from the absence of a feature. Becker (2024) escapes this dilemma by adopting a Word & Paradigm perspective. She avoids morphemic segmentation altogether, and instead focuses on identifying stems automatically (following Beniamine and Guzmán Naranjo 2021; Bonami and Beniamine 2021, with some adjustments for stem allomorphy). She then defines zero-marked forms

as those which consist solely of the stem. Similarly, features are not segmented, and zero markers are considered to mark the entire bundle of morpho-syntactic features for the form. To further reduce potential unfounded proliferation of zero marking, the study employs the perspective of *morphomic paradigms* (Boyé and Schalchi 2016), where any fully syncretic cells in the lexicon are merged.

Becker (2024) finds that overall, zero marking is uncommon. Yet, she observes a lot of variation across languages. Careful statistical analysis reveals this variation to be largely idiosyncrastic. A few trends emerge however: zero-marking is avoided in cells with many values; adjectives and verbs are more likely than nouns to avoid zero marking altogether. Some feature values are comparatively more likely to be zero marked across languages: IMP, SG, 3 and PRS in verbs, NOM, SG and INDF in nouns, NOM.SG in adjectives. Using the Universal Dependency corpora (Zeman *et al.* 2023) to gather frequency information, Becker (2024) confirms the Zipfian effect of frequency on length of overt markers, and finds the effect more pronounced on suffixes than other affixes. Nevertheless, this association does not hold for zero markers, which simply do not behave like short markers. Instead, she confirms the observation from Guzmán Naranjo and Becker 2021 according to which zero markers are dispreferred. This indicates that zero markers may not solely result from phonetic reduction. An alternative path to zero marking more in line with these results would be for them to arise as a distinct, contrastive strategy.

*Guzmán Naranjo: An analogical approach
to the typology of inflectional complexity*

2.3

Guzmán Naranjo (2024) addresses the same conjecture as Cotterell *et al.* (2019) with a new predictive mechanism and at much larger scale. Guzmán Naranjo's model is based on explicit local segmentations of string pairs with variables. Local segmentation is both relatively fast and can be run on very small datasets, since each pair of forms produces a single pattern. Thus, while Cotterell *et al.* require paradigms for at least 700 lexemes to use their neural network method, Guzmán Naranjo is able to analyze on datasets of only 200. Moreover, results from 200-lexeme datasets serve as relatively reliable

lower bounds on the values for larger samples, indicating that even small sets of words can yield useful information about a language.

Guzmán Naranjo (2024) concludes that Cotterell *et al.*'s results do not hold across a larger sample of 71 languages. Although there appears to be a trend relating number of paradigm cells to interpretability, there is no significant correlation. Moreover, he argues that the most valuable measurement of E-complexity is not the number of paradigm cells, but the formal complexity of the rules used to describe them. This sort of E-complexity actually increases as predictability decreases (that is, languages with more complex paradigms are *easier* to predict).

2.4 *Haley et al: Corpus-based measures discriminate inflection and derivation cross-linguistically*

Haley *et al.* (2024) tackle another theoretical question, the division between inflection and derivation. Again, this distinction is the subject of theoretical controversy – Plank (1994) argues that the distinction is gradient rather than categorical, and Haspelmath (2024) claims that it is merely an artifact of traditional linguistic analysis, rather than a phenomenon with real explanatory power. Haley *et al.* propose to characterise morphological relationships by comparing the difference in orthographic form (edit distance) between the related forms, and the difference in corpus distribution (based on FasTex embeddings (Bojanowski *et al.* 2017)), as well as the variability in these measurements across lexemes. Again, while Plank (1994) is able to apply his measures to only 6 morphological relationships, all in English, Haley *et al.* can scale their analysis further, to a set of 26 languages.

Haley *et al.* find that these measurements can be used to predict the traditional divisions between inflection and derivation with relatively high accuracy (variability being more important than magnitude and distribution more important than form). The measurements can also be used to automatically categorise particular constructions as more or less canonically inflectional by ranking their distance to the decision boundary – comparatives, for example, form an intermediate class.

CHALLENGES AND FUTURE DIRECTIONS

3

Yet, the current generation of databases has not made it trivial to run morphological analyses at scale. One set of issues is evident in a comparison between Guzmán Naranjo's 71 languages and Haley *et al.*'s 26, most of which come from Europe: the size of available lexical databases is still closely linked to the kind of information desired. While Unimorph collects inflectional paradigms for a large number of languages, derivational relationships are accessible for far fewer, and corpus embeddings (which have to be collected separately) only for a subset of these. More broadly, there is tension between depth of analysis and typological coverage. The more information is needed, the more the analyst must fall back on scarcer resources which tend to push toward a familiar set of well-resourced European languages.

The interface in the other direction (morphophonology) is similarly problematic. Most available databases list orthographic forms gleaned from dictionaries, but these can preserve antiquated relationships, as in modern French (Baroni 2011), or obscure phonologically predictable ones. Grapheme-to-phoneme conversion is a possible solution, as in Becker 2024 and Mortensen *et al.* 2018, but again, requires resources which may not be available across a typologically diverse sample.

A final issue for lexical databases is the quality and systematicity of the data itself. Gorman *et al.* (2019) register a number of complaints about the quality of the scraped Wiktionary data underlying most Unimorph paradigm tables, including mislabeled cells and misparsed orthographic sequences. Other issues of language in use, such as overabundance (Thornton 2019) and dialectal diversity, can also lead to inconsistencies. While modelling studies like Haley *et al.* 2024 are intended to make analytical categories like 'inflection' and 'derivation' more rigorous by providing more objective ways to make the distinction, the authors acknowledge that this is to some extent undercut by the differing ways in which the database represents purported inflections and derivations in the first place. Similarly, Guzmán Naranjo's decision to include all cliticised and periphrastic forms from Unimorph within his analysis raises theoretical questions of what a word is, or whether such a notion is even cross-linguistically applicable (Dixon

Table 1:
Supplementary
materials

Contribution	Data and code
Inman <i>et al.</i> 2024	https://osf.io/n67mq
Becker 2024	https://osf.io/p4mkc/?view_only=5238ace9cb1d4f4d998486ebb28f4fd8
Guzmán Naranjo 2024	https://doi.org/10.5281/zenodo.11147171
Haley <i>et al.</i> 2024	https://osf.io/uztgy

et al. 2002). In practice, different Unimorph languages make different decisions on what to include within a lexical entry, and this in turn has implications for the rules produced by alignment systems.

Computational approaches to morphological typology greatly benefit from following the FAIR principles (Wilkinson *et al.* 2016), as well as those of Open Science. As shown in Table 1, each contribution in this volume makes their code and data available through open science platforms, in order to facilitate reuse and reproducibility.

Each of the papers in this volume engages with the linguistic literature by testing or sharpening earlier conjectures with reference to newer and larger datasets. In each case, although the authors’ own analysis of their data makes valuable contributions, the work is primarily intended to provide resources (datasets and methods) for future investigation. We hope that the continuing trend of standardization and openness will make large-scale morphological typology more accessible to others within the field, enabling more and more hypotheses to be tested at scale.

REFERENCES

Farrell ACKERMAN and Robert MALOUF (2013), Morphological organization: The low conditional entropy conjecture, *Language*, 89:429–464.

Antonio BARONI (2011), Alphabetic vs. non-alphabetic writing: Linguistic fit and natural tendencies, *The Italian Journal of Linguistics*, 23:127–160, <https://api.semanticscholar.org/CorpusID:56412403>.

Khuyagbaatar BATSUREN, Omer GOLDMAN, Salam KHALIFA, Nizar HABASH, Witold KIERAŚ, Gábor BELLA, Brian LEONARD, Garrett NICOLAI, Kyle GORMAN, Yustinus Ghanggo ATE, Maria RYSKINA, Sabrina MIELKE, Elena BUDIANSKAYA, Charbel EL-KHAISSI, Tiago PIMENTEL, Michael GASSER, William Abbott LANE, Mohit RAJ, Matt COLER, Jaime Rafael Montoya SAMAME, Delio Siticonatzi CAMAITERI, Esaú Zumaeta ROJAS, Didier LÓPEZ FRANCIS, Arturo ONCEVAY, Juan LÓPEZ BAUTISTA, Gema Celeste Silva VILLEGAS, Lucas Torroba HENNIGEN, Adam EK, David GURIEL, Peter DIRIX, Jean-Philippe BERNARDY, Andrey SCHERBAKOV, Aziyana BAYYR-OOL, Antonios ANASTASOPOULOS, Roberto ZARIQUIEY, Karina SHEIFER, Sofya GANIEVA, Hilaria CRUZ, Ritván KARAHÓĞA, Stella MARKANTONATOU, George PAVLIDIS, Matvey PLUGARYOV, Elena KLYACHKO, Ali SALEHI, Candy ANGULO, Jatayu BAXI, Andrew KRIZHANOVSKY, Natalia KRIZHANOVSKAYA, Elizabeth SALESKY, Clara VANIA, Sardana IVANOVA, Jennifer WHITE, Rowan Hall MAUDSLAY, Josef VALVODA, Ran ZMIGROD, Paula CZARNOWSKA, Irene NIKKARINEN, Aelita SALCHAK, Brijesh BHATT, Christopher STRAUGHN, Zoey LIU, Jonathan North WASHINGTON, Yuval PINTER, Duygu ATAMAN, Marcin WOLIŃSKI, Totok SUHARDIJANTO, Anna YABLONSKAYA, Niklas STOHR, Hossep DOLATIAN, Zahroh NURIAH, Shyam RATAN, Francis M. TYERS, Edoardo M. PONTI, Grant AITON, Aryaman ARORA, Richard J. HATCHER, Ritesh KUMAR, Jeremiah YOUNG, Daria RODIONOVA, Anastasia YEMELINA, Taras ANDRUSHKO, Igor MARCHENKO, Polina MASHKOVTSOVA, Alexandra SEROVA, Emily PRUD'HOMMEAUX, Maria NEPOMNIASHCHAYA, Fausto GIUNCHIGLIA, Eleanor CHODROFF, Mans HULDEN, Miikka SILFVERBERG, Arya D. MCCARTHY, David YAROWSKY, Ryan COTTERELL, Reut TSARFATY, and Ekaterina VYLOMOVA (2022), UniMorph 4.0: Universal Morphology, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 840–855, European Language Resources Association, Marseille, France, <https://aclanthology.org/2022.lrec-1.89>.

Laura BECKER (2024), Zero marking in inflection: A token-based approach, in *Computational Approaches to Morphological Typology*, TODO.

Sacha BENIAMINE, Cormac ANDERSON, Mae CARROLL, Matías Guzmán NARANJO, Borja HERCE, Matteo PELLEGRINI, Erich ROUND, Helen SIMS-WILLIAMS, and Tiago TRESOLDI (2023), Paralex: a DeAR standard for rich lexicons of inflected forms, in *Presentation at International Symposium of Morphology*, https://ismo2023.ovh/fichiers/abstracts/4_ISMO_2023_Paralex.pdf, <https://www.paralex-standard.org>.

Sacha BENIAMINE and Matías GUZMÁN NARANJO (2021), Multiple alignments of inflectional paradigms, *Proceedings of the Society for Computation in Linguistics*, 4:216–227.

- Balthasar BICKEL and Johanna NICHOLS (2002), Autotypologizing databases and their use in fieldwork, in *Proceedings of the international LREC workshop on resources and tools in field linguistics, Las Palmas*, volume 2627, MPI for Psycholinguistics Nijmegen, doi:10.5167/UZH-76860.
- Balthasar BICKEL, Johanna NICHOLS, Taras ZAKHARKO, Alena WITZLACK-MAKAREVICH, Kristine HILDEBRANDT, Michael RIESSLER, Lennart BIERKANDT, Fernando ZÚÑIGA, and John B. LOWE (2022), The autotyp database, doi:10.5281/ZENODO.6793367.
- Piotr BOJANOWSKI, Edouard GRAVE, Armand JOULIN, and Tomas MIKOLOV (2017), Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Olivier BONAMI and Sacha BENIAMINE (2021), Leaving the stem by itself, in Sedigheh MORADI, Marcia HAAG, Janie REES-MILLER, and Andrija PETROVIC, editors, *All things morphology: Its independence and its interfaces*, pp. 81–98, Benjamins, doi:10.1075/cilt.353.05bon.
- Gilles BOYÉ and Gauvin SCHALCHI (2016), The status of paradigms, in Andrew HIPPLEY and Gregory STUMP, editors, *The Cambridge handbook of morphology*, pp. 206–234, Cambridge University Press.
- Dunstan BROWN, Marina CHUMAKINA, and Greville G. CORBETT (2012), *Canonical morphology and syntax*, Oxford University Press, ISBN 9780199604326, doi:10.1093/acprof:oso/9780199604326.001.0001, <https://doi.org/10.1093/acprof:oso/9780199604326.001.0001>.
- Thea CAMERON-FAULKNER and Andrew CARSTAIRS-MCCARTHY (2000), Stem alternants as morphological signata: Evidence from blur avoidance in Polish nouns, *Natural Language and Linguistic Theory*, 18:813–835.
- A. D. CARSTAIRS (1987), *Allomorphy in inflexion*, Croom Helm linguistics series, Croom Helm, London.
- Greville G. CORBETT (2005), *The canonical approach in typology*, pp. 25–49, John Benjamins Publishing Company, doi:10.1075/slcs.72.03cor.
- Greville G. CORBETT (2023), The typology of external splits, 99(1):108–153, ISSN 1535-0665, doi:10.1353/lan.2023.0007.
- Ryan COTTERELL, Christo KIROV, Mans HULDEN, and Jason EISNER (2019), On the complexity and typology of inflectional morphological systems, *Transactions of the Association for Computational Linguistics*, 7:327–342.
- Robert MW DIXON, Alexandra Y AIKHENVALD, et al. (2002), *Word: A cross-linguistic typology*, Cambridge University Press.
- Kyle GORMAN, Arya D. MCCARTHY, Ryan COTTERELL, Ekaterina VYLOMOVA, Miikka SILFVERBERG, and Magdalena MARKOWSKA (2019), Weird inflects but OK: Making sense of morphological generation errors, in Mohit BANSAL and

Aline VILLAVICENCIO, editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 140–151, Association for Computational Linguistics, Hong Kong, China, doi:10.18653/v1/K19-1014, <https://aclanthology.org/K19-1014>.

Joseph GREENBERG (1966), *Language universals*, number 59 in *Janua Linguarum. Series Minor*, De Gruyter Mouton, 3rd printing. reprint 2019 edition, ISBN 9783110802528.

Joseph H. GREENBERG (1960), A quantitative approach to the morphological typology of language, *International Journal of American Linguistics*, 26(3):178–184.

Matías GUZMÁN NARANJO (2024), An analogical approach to the typology of inflectional complexity, in *Computational Approaches to Morphological Typology*, TODO.

Matías GUZMÁN NARANJO and Laura BECKER (2021), Coding efficiency in nominal inflection: Expectedness and type frequency effects, *Linguistics Vanguard*, 7(s3):20190075, doi:10.1515/lingvan-2019-0075.

Coleman HALEY, Eduardo M. PONTI, and Sharon GOLDWATER (2024), Corpus-based measures discriminate inflection and derivation cross-linguistically, in *Computational Approaches to Morphological Typology*, TODO.

Martin HASPELMATH (2008), *8 creating economical morphosyntactic patterns in language change*, pp. 185–214, Oxford University Press Oxford, ISBN 9780191711442, doi:10.1093/acprof:oso/9780199298495.003.0008.

Martin HASPELMATH (2024), Inflection and derivation as traditional comparative concepts, *Linguistics*, 62(1):43–77.

Martin HASPELMATH, Matthew S. DRYER, David GIL, and Bernard COMRIE (2013), *The world atlas of language structures online*, Max Planck Digital Library, <http://wals.info>.

David INMAN, Alena WITZLACK-MAKAREVICH, Natalia CHOUSOU-POLYDOURI, and Melvin STEIGER (2024), Alignment everywhere all at once: Applying the late aggregation principle to a typological database of argument marking, in *Computational Approaches to Morphological Typology*, TODO.

Christo KIROV, Ryan COTTERELL, John SYLAK-GLASSMAN, Géraldine WALTHER, Ekaterina VYLOMOVA, Patrick XIA, Manaal FARUQUI, Sabrina J. MIELKE, Arya MCCARTHY, Sandra KÜBLER, David YAROWSKY, Jason EISNER, and Mans HULDEN (2018), UniMorph 2.0: Universal Morphology, in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan, <https://aclanthology.org/L18-1293>.

Christo KIROV, John SYLAK-GLASSMAN, Roger QUE, and David YAROWSKY (2016), Very-large scale parsing and normalization of wiktionary morphological paradigms, in Nicoletta Calzolari (Conference CHAIR), Khalid CHOUKRI, Thierry DECLERCK, Sara GOGGI, Marko GROBELNIK, Bente MAEGAARD, Joseph MARIANI, Helene MAZO, Asuncion MORENO, Jan ODIJK, and Stelios PIPERIDIS, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), Paris, France, ISBN 978-2-9517408-9-1, <http://ckirov.github.io/UniMorph/>.

Arya D. MCCARTHY, Christo KIROV, Matteo GRELLA, Amrit NIDHI, Patrick XIA, Kyle GORMAN, Ekaterina VYLOMOVA, Sabrina J. MIELKE, Garrett NICOLAI, Miikka SILFVERBERG, Timofey ARKHANGELSKIY, Nataly KRIZHANOVSKY, Andrew KRIZHANOVSKY, Elena KLYACHKO, Alexey SOROKIN, John MANSFIELD, Valts ERNŠTREITS, Yuval PINTER, Cassandra L. JACOBS, Ryan COTTERELL, Mans HULDEN, and David YAROWSKY (2020), UniMorph 3.0: Universal Morphology, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3922–3931, European Language Resources Association, Marseille, France, ISBN 979-10-95546-34-4, <https://aclanthology.org/2020.lrec-1.483>.

David R. MORTENSEN, Siddharth DALMIA, and Patrick LITTELL (2018), Epitran: Precision G2P for many languages, in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Frans PLANK (1994), Inflection and derivation, in *The Encyclopedia of Language and Linguistics*, pp. 1671–1679, Elsevier Science and Technology.

Hedvig SKIRGÅRD, Hannah J. HAYNIE, Damián E. BLASI, Harald HAMMARSTRÖM, Jeremy COLLINS, Jay J. LATARCHE, Jakob LESAGE, Tobias WEBER, Alena WITZLACK-MAKAREVICH, Sam PASSMORE, *et al.* (2023), Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss, *Science Advances*, 9(16):eadg6175.

Gregory T. STUMP and Raphael FINKEL (2013), *Morphological typology: From word to paradigm*, Cambridge University Press.

Anna M. THORNTON (2019), Overabundance: A canonical typology, *Competition in inflection and word-formation*, pp. 223–258.

A.W. VON SCHLEGEL (1818), *Observations sur la langue et la littérature provençales*, Librairie grecque-latine-allemande.

Mark D. WILKINSON, Michel DUMONTIER, IJsbrand Jan AALBERSBERG, Gabrielle APPLETON, Myles AXTON, Arie BAAK, Niklas BLOMBERG, Jan-Willem BOITEN, Luiz Bonino DA SILVA SANTOS, Philip E. BOURNE, Jildau BOUWMAN, Anthony J. BROOKES, Tim CLARK, Mercè CROSAS, Ingrid DILLO, Olivier DUMON, Scott EDMUNDS, Chris T. EVELO, Richard FINKERS, Alejandra

GONZALEZ-BELTRAN, Alasdair J. G. GRAY, Paul GROTH, Carole GOBLE, Jeffrey S. GRETHE, Jaap HERINGA, Peter A. C. 'T HOEN, Rob HOOFT, Tobias KUHN, Ruben KOK, Joost KOK, Scott J. LUSHER, Maryann E. MARTONE, Albert MONS, Abel L. PACKER, Bengt PERSSON, Philippe ROCCA-SERRA, Marco ROOS, Rene VAN SCHAIK, Susanna-Assunta SANSONE, Erik SCHULTES, Thierry SENGSTAG, Ted SLATER, George STRAWN, Morris A. SWERTZ, Mark THOMPSON, Johan VAN DER LEI, Erik VAN MULLIGEN, Jan VELTEROP, Andra WAAGMEESTER, Peter WITTENBURG, Katherine WOLSTENCROFT, Jun ZHAO, and Barend MONS (2016), The fair guiding principles for scientific data management and stewardship, *Scientific Data*, 3(1):160018, ISSN 2052-4463, doi:10.1038/sdata.2016.18.

Alena WITZLACK-MAKAREVICH, Johanna NICHOLS, Kristine A. HILDEBRANDT, Taras ZAKHARKO, and Balthasar BICKEL (2022), *Managing autotyp data: Design principles and implementation*, pp. 631–642, The MIT Press, ISBN 9780262366076, doi:10.7551/mitpress/12200.003.0061.

Daniel ZEMAN, Joakim NIVRE, Mitchell ABRAMS, Elia ACKERMANN, Noëmi AEPLI, Hamid AGHAEI, Željko AGIĆ, Amir AHMADI, Lars AHRENBERG, Chika Kennedy AJEDE, Gabrielė ALEKSANDRAVIČIŪTĖ, Ika ALFINA, Lene ANTONSEN, Katya APLONOVA, Angelina AQUINO, Carolina ARAGON, Maria Jesus ARANZABE, Bilge Nas ARICAN, Órunn ARNARDÓTTIR, Gashaw ARUTIE, Jessica Naraiswari ARWIDARASTI, Masayuki ASAHARA, Deniz Baran ASLAN, Luma ATEYAH, Furkan ATMACA, Mohammed ATTIA, Aitziber ATUTXA, Liesbeth AUGUSTINUS, Elena BADMAEVA, Keerthana BALASUBRAMANI, Miguel BALLESTEROS, Esha BANERJEE, Sebastian BANK, Verginica BARBU MITITELU, Starkađur BARKARSON, Rodolfo BASILE, Victoria BASMOV, Colin BATCHELOR, John BAUER, Seyyit Talha BEDIR, Kepa BENGOTXEA, Gözde BERK, Yevgeni BERZAK, Irshad Ahmad BHAT, Riyaz Ahmad BHAT, Erica BIAGETTI, Eckhard BICK, Agnė BIELINSKIENĖ, Kristín BJARNADÓTTIR, Rogier BLOKLAND, Victoria BOBICEV, Loïc BOIZOU, Emanuel BORGES VÖLKER, Carl BÖRSTELL, Cristina BOSCO, Gosse BOUMA, Sam BOWMAN, Adriane BOYD, Anouck BRAGGAAR, Kristina BROKAITĖ, Aljoscha BURCHARDT, Marie CANDITO, Bernard CARON, Gauthier CARON, Lauren CASSIDY, Tatiana CAVALCANTI, Gülşen CEBIROĞLU ERYİĞİT, Flavio Massimiliano CECCHINI, Giuseppe G. A. CELANO, Slavomír ČEPLÖ, Neslihan CESUR, Savas CETIN, Özlem ÇETİNOĞLU, Fabricio CHALUB, Shweta CHAUHAN, Ethan CHI, Taishi CHIKA, Yongseok CHO, Jinho CHOI, Jayeol CHUN, Juyeon CHUNG, Alessandra T. CIGNARELLA, Silvie CINKOVÁ, Aurélie COLLOMB, Çağrı ÇÖLTEKİN, Miriam CONNOR, Marine COURTIN, Mihaela CRISTESCU, Philemon DANIEL, Elizabeth DAVIDSON, Marie-Catherine DE MARNEFFE, Valeria DE PAIVA, Mehmet Oguz DERIN, Elvis DE SOUZA, Arantza DIAZ DE ILARRAZA, Carly DICKERSON, Arawinda DINAKARAMANI, Elisa DI NUOVO, Bamba DIONE, Peter DIRIX, Kaja DOBROVOLJC, Timothy DOZAT, Kira DROGANOVA, Puneet DWIVEDI, Hanne ECKHOFF, Sandra EICHE, Marhaba ELI, Ali ELKAHKY,

Binyam EPHREM, Olga ERINA, Tomaž ERJAVEC, Aline ETIENNE, Wograinne EVELYN, Sidney FACUNDES, Richárd FARKAS, Jannatul FERDAOUSI, Marília FERNANDA, Hector FERNANDEZ ALCALDE, Jennifer FOSTER, Cláudia FREITAS, Kazunori FUJITA, Katarína GAJDOŠOVÁ, Daniel GALBRAITH, Marcos GARCIA, Moa GÄRDENFORS, Sebastian GARZA, Fabrício Ferraz GERARDI, Kim GERDES, Filip GINTER, Gustavo GODOY, Iakes GOENAGA, Koldo GOJENOLA, Memduh GÖKIRMAK, Yoav GOLDBERG, Xavier GÓMEZ GUINOVART, Berta GONZÁLEZ SAAVEDRA, Bernadeta GRICIŪTĖ, Matias GRIONI, Loïc GROBOL, Normunds GRŪZTIS, Bruno GUILLAUME, Céline GUILLOT-BARBANCE, Tunga GÜNGÖR, Nizar HABASH, Hinrik HAFSTEINSSON, Jan HAJIČ, Jan HAJIČ JR., Mika HÄMÄLÄINEN, Linh HÀ MỸ, Na-Rae HAN, Muhammad Yudistira HANIFMUTI, Sam HARDWICK, Kim HARRIS, Dag HAUG, Johannes HEINECKE, Oliver HELLWIG, Felix HENNIG, Barbora HLADKÁ, Jaroslava HLAVÁČOVÁ, Florinel HOCIUNG, Petter HOHLE, Eva HUBER, Jena HWANG, Takumi IKEDA, Anton Karl INGASON, Radu ION, Elena IRIMIA, Ȑlájídė ISHOLA, Kaoru ITO, Siratun JANNAT, Tomáš JELÍNEK, Apoorva JHA, Anders JOHANNSEN, Hildur JÓNSDÓTTIR, Fredrik JØRGENSEN, Markus JUUTINEN, Sarveswaran K., Hüner KAŞIKARA, Andre KAASEN, Nadezhda KABAEVA, Sylvain KAHANE, Hiroshi KANAYAMA, Jenna KANERVA, Neslihan KARA, Boris KATZ, Tolga KAYADELEN, Jessica KENNEY, Václava KETTNEROVÁ, Jesse KIRCHNER, Elena KLEMENTIEVA, Elena KLYACHKO, Arne KÖHN, Abdullatif KÖKSAL, Kamil KOPACEWICZ, Timo KORKIAKANGAS, Mehmet KÖSE, Natalia KOTSYBA, Jolanta KOVALEVSKAITĖ, Simon KREK, Parameswari KRISHNAMURTHY, Sandra KÜBLER, Oğuzhan KUYRUKÇU, Asli KUZGUN, Sookyoung KWAK, Veronika LAIPPALA, Lucia LAM, Lorenzo LAMBERTINO, Tatiana LANDO, Septina Dian LARASATI, Alexei LAVRENTIEV, John LEE, Phươg Lê HỒNG, Alessandro LENCI, Saran LERTPRADIT, Herman LEUNG, Maria LEVINA, Cheuk Ying LI, Josie LI, Keying LI, Yuan LI, KyungTae LIM, Bruna LIMA PADOVANI, Krister LINDÉN, Nikola LJUBEŠIĆ, Olga LOGINOVA, Stefano LUSITO, Andry LUTHFI, Mikko LUUKKO, Olga LYASHEVSKAYA, Teresa LYNN, Vivien MACKETANZ, Menel MAHAMDI, Jean MAILLARD, Aibek MAKAZHANOV, Michael MANDL, Christopher MANNING, Ruli MANURUNG, Büşra MARŞAN, Cătălina MĂRÂNDUC, David MAREČEK, Katrin MARHEINECKE, Héctor MARTÍNEZ ALONSO, Lorena MARTÍN-RODRÍGUEZ, André MARTINS, Jan MAŠEK, Hiroshi MATSUDA, Yuji MATSUMOTO, Alessandro MAZZEI, Ryan McDONALD, Sarah McGUINNESS, Gustavo MENDONÇA, Tatiana MERZHEVICH, Niko MIEKKA, Karína MISCHENKOVA, Margarita MISIRPASHAYEVA, Anna MISSILÄ, Cătălin MITITELU, Maria MITROFAN, Yusuke MIYAO, AmirHossein MOJIRI FOROUSHANI, Judit MOLNÁR, Amirsaeid MOLOODI, Simonetta MONTEMAGNI, Amir MORE, Laura MORENO ROMERO, Giovanni MORETTI, Keiko Sophie MORI, Shinsuke MORI, Tomohiko MORIOKA, Shigeki MORO, Bjartur MORTENSEN, Bohdan MOSKALEVSKYI, Kadri MUISCHNEK, Robert MUNRO, Yugo MURAWAKI, Kaili MÜÜRISep, Pinkey

NAINWANI, Mariam NAKHLÉ, Juan Ignacio NAVARRO HORÑIACEK, Anna NEDOLUZHKO, Gunta NEŠPORE-BĚRZKALNE, Manuela NEVACI, Lương NGUYỄN THỊ, Huyền NGUYỄN THỊ MINH, Yoshihiro NIKAIDO, Vitaly NIKOLAEV, Rattima NITISAROJ, Alireza NOURIAN, Hanna NURMI, Stina OJALA, Atul Kr. OJHA, Adédayo OLÚÒKUN, Mai OMURA, Emeka ONWUEGBUZIA, Petya OSENOVA, Robert ÖSTLING, Lilja ØVRELID, Şaziye Betül ÖZATEŞ, Merve ÖZÇELİK, Arzucan ÖZGÜR, Balkız ÖZTÜRK BAŞARAN, Hyunji Hayley PARK, Niko PARTANEN, Elena PASCUAL, Marco PASSAROTTI, Agnieszka PATEJUK, Guilherme PAULINO-PASSOS, Angelika PELJAK-ŁAPIŃSKA, Siyao PENG, Cenel-Augusto PEREZ, Natalia PERKOVA, Guy PERRIER, Slav PETROV, Daria PETROVA, Jason PHELAN, Jussi PIITULAINEN, Tommi A PIRINEN, Emily PITLER, Barbara PLANK, Thierry POIBEAU, Larisa PONOMAREVA, Martin POPEL, Lauma PRETKALNIŅA, Sophie PRÉVOST, Prokopis PROKOPIDIS, Adam PRZEPIÓRKOWSKI, Tiina PUOLAKAINEN, Sampo PYYSALO, Peng QI, Andriela RÄÄBIS, Alexandre RADEMAKER, Mizanur RAHOMAN, Taraka RAMA, Loganathan RAMASAMY, Carlos RAMISCH, Fam RASHEL, Mohammad Sadegh RASOOLI, Vinit RAVISHANKAR, Livy REAL, Petru REBEJA, Siva REDDY, Mathilde REGNAULT, Georg REHM, Ivan RIABOV, Michael RIESSLER, Erika RIMKUTĖ, Larissa RINALDI, Laura RITUMA, Putri RIZQIYAH, Luisa ROCHA, Eiríkur RÖGNVALDSSON, Mykhailo ROMANENKO, Rudolf ROSA, Valentin ROŞCA, Davide ROVATI, Olga RUDINA, Jack RUETER, Kristján RÚNARSSON, Shoal SADDE, Pegah SAFARI, Benoît SAGOT, Aleksī SAHALA, Shadi SALEH, Alessio SALOMONI, Tanja SAMARDŽIĆ, Stephanie SAMSON, Manuela SANGUINETTI, Ezgi SANIYAR, Dage SÄRG, Baiba SAULTE, Yanin SAWANAKUNANON, Shefali SAXENA, Kevin SCANNELL, Salvatore SCARLATA, Nathan SCHNEIDER, Sebastian SCHUSTER, Lane SCHWARTZ, Djamé SEDDAH, Wolfgang SEEKER, Mojgan SERAJI, Syeda SHAHZADI, Mo SHEN, Atsuko SHIMADA, Hiroyuki SHIRASU, Yana SHISHKINA, Muh SHOHIBUSSIRRI, Dmitry SICHINAVA, Janine SIEWERT, Einar Freyr SIGURÐSSON, Aline SILVEIRA, Natalia SILVEIRA, Maria SIMI, Radu SIMIONESCU, Katalin SIMKÓ, Mária ŠIMKOVÁ, Kiril SIMOV, Maria SKACHEDUBOVA, Aaron SMITH, Isabela SOARES-BASTOS, Shafi SOUROV, Carolyn SPADINE, Rachele SPRUGNOLI, Steinór STEINGRÍMSSON, Antonio STELLA, Milan STRAKA, Emmett STRICKLAND, Jana STRNADOVÁ, Alane SUHR, Yogi Lesmana SULESTIO, Umut SULUBACAK, Shingo SUZUKI, Zsolt SZÁNTÓ, Chihiro TAGUCHI, Dima TAJI, Yuta TAKAHASHI, Fabio TAMBURINI, Mary Ann C. TAN, Takaaki TANAKA, Dipta TANAYA, Samson TELLA, Isabelle TELLIER, Marinella TESTORI, Guillaume THOMAS, Liisi TORG, Marsida TOSKA, Trond TROSTERUD, Anna TRUKHINA, Reut TSARFATY, Utku TÜRK, Francis TYERS, Sumire UEMATSU, Roman UNTILOV, Zdeňka UREŠOVÁ, Larraitx URÍA, Hans USZKOREIT, Andrius UTKA, Sowmya VAJJALA, Rob VAN DER GOOT, Martine VANHOVE, Daniel VAN NIEKERK, Gertjan VAN NOORD, Viktor VARGA, Eric VILLEMONT DE LA CLERGERIE, Veronika VINCZE, Natalia VLASOVA, Aya WAKASA, Joel C.

WALLENBERG, Lars WALLIN, Abigail WALSH, Jing Xian WANG, Jonathan North WASHINGTON, Maximilan WENDT, Paul WIDMER, Sri Hartati WIJONO, Seyi WILLIAMS, Mats WIRÉN, Christian WITTERN, Tsegay WOLDEMARIAM, Tak-sum WONG, Alina WRÓBLEWSKA, Mary YAKO, Kayo YAMASHITA, Naoki YAMAZAKI, Chunxiao YAN, Koichi YASUOKA, Marat M. YAVRUMYAN, Arife Betül YENICE, Olcay Taner YILDIZ, Zhuoran YU, Arlisa YULIAWATI, Zdeněk ŽABOKRTSKÝ, Shorouq ZAHRA, Amir ZELDES, He ZHOU, Hanzhi ZHU, Anna ZHURAVLEVA, and Rayan ZIANE (2023), Universal Dependencies 2.13, <http://hdl.handle.net/11234/1-5287>.

George Kingsley ZIPF (2013), *The psycho-biology of language*, Routledge, ISBN 9781136310461, doi:10.4324/9781315009421.

Micha Elsner

Ⓘ 0000-0002-1432-2129

elsner.14@osu.edu

Ohio State University

Sacha Beniamine

Ⓘ 0000-0003-2584-3576

s.beniamine@surrey.ac.uk

University of Surrey

Micha Elsner and Sacha Beniamine (2024), *Computational approaches to morphological typology*, Journal of Language Modelling, 12(2):271–286

Ⓓ <https://dx.doi.org/10.15398/jlm.v12i2.431>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

ⒸⒻ <http://creativecommons.org/licenses/by/4.0/>

Alignment everywhere all at once: Applying the late aggregation principle to a typological database of argument marking

David Inman^{1*}, Alena Witzlack-Makarevich^{2*},
Natalia Chousou-Polydouri¹, and Melvin Steiger^{1*}

¹ University of Zurich & Center for the Interdisciplinary Study of Language
Evolution

² Hebrew University of Jerusalem

ABSTRACT

This article presents the structure of the ATLAs Alignment Module, a typological database designed to exhaustively capture language-internal variation in argument marking (indexing and flagging). The flexible design of our database can be extended to cover further aspects of morphosyntactic alignment. We demonstrate with a small diversity sample how the database can be queried and the data aggregated at different levels of structure (e.g. for a language as a whole or for individual referential types in the form of alignment statements) for the purposes of cross-linguistic comparison. The database is made available in the Cross-Linguistic Data Formats (CLDF), and we provide code that generates an array of aggregations.

Keywords:
alignment,
typology,
database,
aggregation,
morphology

INTRODUCTION

1

Alignment of argument marking is one of the major morphosyntactic characteristics of languages both in the descriptions of individual languages as well as in comparative studies and typological databases.

*These authors have contributed equally to this work.

All major typological databases, such as WALS (Dryer and Haspelmath 2013) and Grambank (Skirgård *et al.* 2023), include several alignment-related features. Furthermore, a dedicated database tracks the emergence of alignment patterns (Cristofaro *et al.* 2021).¹ Typological work on morphosyntactic alignment (including the aforementioned databases) typically captures only high-level generalizations about alignment at the level of the entire language, e.g. the presence of ergativity in the system of case marking or traces of hierarchical effects on the agreement system. However, many languages have multiple alignments conditioned e.g. by the referential properties of arguments, the tense of the clause, and so on (see e.g. Bickel *et al.* 2015b). Thus, in contrast to some other typological features (e.g. presence of a nominal dual number), alignment is not a typological variable for which there is only one way to make a statement about a language as a whole. Instead, it is a complex and multi-variate component of grammar for which similarities and differences between languages can be established along many different dimensions.

In this article we present the ATLAs Alignment Module (Inman *et al.* in prep), a typological database of morphosyntactic alignment designed to capture existing variation in alignment patterns of a language. By encoding multiple aspects and patterns of alignment within a language all at once, we will show that it is possible to aggregate alignment information at differing levels of structure: for the language as a whole, for individual argument selectors (e.g. nominative case, plural argument marker), for individual referential types (e.g. 1sg, 2pl, masculine nouns), and for argument roles (S, A, and P).

We will begin with an overview of the phenomenon of alignment (Section 2) and discuss how data that describe the phenomenon can be captured in typological databases (Section 3). We will then describe the choices we made for data collection and database design (Section 4) and demonstrate how the data we collected can be used to derive a variety of typological properties (Section 5). Finally, we will offer some concluding remarks and discuss the ways in which our work can

¹ Alignment-related information is also captured in databases dedicated to valency patterns (Hartmann *et al.* 2013, Say 2020–). Note, however, that these databases focus on predicate-level details and variations of predicate-specific coding frames.

be extended to answer more questions about alignment (Section 6). Supplementary Materials, including the full database and all code, are available at <https://osf.io/n67mq/>.

MORPHOSYNTACTIC ALIGNMENT

2

The study of morphosyntactic alignment is intimately linked with the broader phenomenon of *grammatical relations*. This label traditionally refers to the relations between a clause or a predicate and its arguments. Some of the most common grammatical relations are subject and object, which are among the basic concepts of many theoretical frameworks. However, starting from the mid-1970s, descriptive linguists and typologists have reported challenges in identifying such traditional grammatical relations in individual languages and in applying them consistently in typological studies (see in particular the collection of papers in Li and Thompson 1976, LaPolla 1993, and Dryer 1996, 1997).

Most typologically informed research adopts a language-specific and construction-specific view of grammatical relations (cf. Comrie 1978; Moravcsik 1978; Van Valin 1981, 1983, 2005; Croft 2001; Bickel 2011; Witzlack-Makarevich 2011, 2019). In this approach, researchers forego assumptions about the universality of grammatical relations, such as subject and object. Instead, they use more robust cross-linguistic concepts as a point of comparison for the relevant morphosyntactic properties of arguments or constructions.² In what follows, we first provide an overview of these concepts.

² A classic early example of the objectors of this approach is Anderson (1976), who argues that the switch reference construction is the only right way to determine what a subject is in the language Kâte [kate1253] (Nuclear Trans New Guinea; Papua New Guinea), which has ergative flagging and accusative indexing (see Section 2.2). This is a case of prioritizing the identification of a specific grammatical relation (in this case, “subject”) over considering all relevant morphosyntactic facts of the language.

2.1

Arguments

A common way to capture how arguments of a clause are treated by various morphosyntactic constructions in individual languages is to ask which arguments are marked or behave in the same way. This identity of marking or behavior of certain arguments is what is understood as morphosyntactic alignment. Consider the case marking of the noun ‘man’ in the Chechen example in (1) and its English translation.

- (1) Chechen [chec1245] (Nakh-Daghestanian; Russia; Zarina Molochieva p.c.)
- a. *stag valla.*
man die.PRF
‘The man died.’
 - b. *stag-as xudar de’a-na.*
man-ERG porridge eat-PRF
‘The man has eaten porridge.’
 - c. *ʒʃala-s stag qieri-na.*
dog-ERG man frighten-PRF
‘The dog frightened the man.’

Whereas the arguments the dead man in (1a), the eating man in (1b) and the frightened man (1c) in the English translation do not have any overt case marking (it is just ‘the man’), the Chechen examples have two types of argument marking: the dead man and the frightened man are not marked in any visible way, while the eating man has the dedicated case suffix *-(a)s*, which linguists commonly refer to as *ergative case suffix*. If you translate these sentences into a language which has a special accusative case, the overall picture will be quite different: the frightened man will be marked in a special way and thus differently from the dead man and the eating man, which would be in the (unmarked) nominative case.

This marking is not a special property of the word ‘man’ and the verbs included in these examples. Instead, it is a pattern found with other nominal and pronominal arguments and other verbs across the language, so we need a way to generalize across arguments of different predicates. As we will outline in this section, we understand arguments

as a composite category made of both *argument role* and *referential properties*. We will first outline how we define argument roles.

The most common argument roles used for the purposes of alignment typology and in descriptive accounts are S, A, and P (or O in some sources).³ Note, however, that what exactly is understood by these labels varies somewhat between authors (see Haspelmath 2011). We use these terms in the sense of *generalized semantic argument roles* (as opposed to a semantic-syntactic or purely syntactic understanding). A generalized semantic argument role (henceforth *argument role* or just *role*) is an abstraction over *predicate-specific roles* (or *microroles*, as e.g. in Hartmann *et al.* 2013). For example, the verb *hit* has two predicate-specific roles, a *HITTER* and a *HITTEE*, the verb *kiss* has a *KISSER* and a *KISSEE*, *see* has a *SEER* and a *SEEN*, and so on. In the case of the role A, it abstracts over the predicate-specific roles of *HITTER*, *KISSER*, and *SEER*, according to semantic criteria we list below.

Argument roles are first distinguished according to the numerical valency of their predicates: the sole argument of one-argument predicates vs. the two arguments of two-argument predicates. In the case of the sole argument of one-argument (monovalent) predicates, there is no need to distinguish it from anything else; this argument is abbreviated as S, independent of its finer semantic differences. In the case of two-argument (bivalent) predicates, arguments are distinguished on the basis of cross-linguistically viable lexical entailment properties (as in Witzlack-Makarevich 2011, 2019, following Dowty 1991 and Primus 1999, 2006).

Each argument of a bivalent verb accumulates various lexical entailment properties, given in (2). The argument that accumulates more lexical entailments than the other argument of the same verb is the A role, and the other is the P role.

- (2) a. causing an event (e.g. A hits P, A kisses P, A goes to P)

³The alignment of other argument types, in particular, of the arguments of trivalent or ditransitive verbs, is another common research topic, see e.g. the collections of papers in Malchukov *et al.* 2010b. Due to the project scope, we do not treat any other argument roles apart from S, A, and P. However, the framework presented in Section 4 is equipped and sufficiently flexible to incorporate other domains of alignment, including the alignment of ditransitive verbs.

- b. volitional (e.g. A hits P, A kisses P)
- c. sentient (e.g. A sees P, A looks at P, A loves P, P pleases A)
- d. independently existing (e.g. A bakes P, A makes P)
- e. possessing another participant (e.g. A has P, P belongs to A)

For instance, in *Lisa kisses Mario*, *Lisa* is causing the event, she is volitional and sentient, and she exists independently. On the other hand, *Mario* only is sentient and exists independently in this event. Thus, *Lisa* accumulates more of the relevant properties than *Mario* and is classified as A. The remaining argument (*Mario*) is P. Thus, every two-argument predicate will have one argument which can be labelled as A and one which can be labelled as P, following the list of lexical entailments in (2). Note that this labeling process is determined entirely by semantics: there is no reference here to syntactic structure or morphological marking.

With this cross-linguistically applicable set of argument roles, it is possible to calculate alignments by comparing the marking or the behavior of different roles. The five logically possible alignment types are listed in (3). We will refer to them as *basic alignment types*.

- (3) Basic alignment types
- a. $S = A \neq P$ corresponds to the (nominative-)accusative alignment pattern (S and A are marked or behave identically but differently from P);
 - b. $S = P \neq A$ corresponds to the ergative(-absolutive) alignment pattern;
 - c. $S = A = P$ corresponds to the neutral alignment pattern;
 - d. $S \neq A \neq P$ corresponds to the tripartite alignment pattern;
 - e. $A = P \neq S$ corresponds to the horizontal alignment pattern.

These five basic alignment patterns figure prominently in many typological studies, both dedicated to alignment specifically (e.g. Comrie 2013a,b; Siewierska 2013a) and in large-scale studies of genealogical, geographic, and universal determinants of linguistic patterning (e.g. Nichols 1992). The list in (3) is often expanded with further non-basic alignment types meant to capture specific patterns of

argument marking. For example, Siewierska (2013a) adds active, hierarchical, and split alignment to the list of possible values.

As we have noted at the beginning of this section, arguments have a composite structure in the approach we adopt (see Bickel 2011): In addition to the argument roles, various referential properties of arguments (person, number, definiteness, topicality, specificity, animacy, and also part of speech) can determine the argument's marking by indexing or flagging and thus have an immediate effect on alignment (as demonstrated in Section 2.3).

Argument selectors

2.2

There are two major ways in which some arguments can be treated identically by a language's grammar: via patterns of morphological marking (also called *coding*, or just *marking*) and via patterns of (syntactic) behavior. Coding traditionally encompasses different loci of morphological marking, both case marking on the noun phrase and indexing on the verb (or in clausal inflection), as well as word order (Keenan 1976). We will refer to all ways in which a language groups arguments, either syntactically or morphologically, as *argument selectors*, and will furthermore focus on morphological marking, leaving aside word order. Cross-linguistically, by far the most common argument selectors, as well as the best studied ones, are *flagging* and *indexing*.⁴

We use the term *flagging*, following Malchukov *et al.* (2010a, 8), as a cover term for both morphological case and adposition marking, both of which mark a role within the syntactic domain of a noun phrase. We use the term *indexing* to refer to the marking of verbal agreement or argument cross-referencing on the clause as a whole (again, following Malchukov *et al.*). The present study only concerns the argument selectors of flagging and indexing.

⁴ The set of syntactic (or *behavioral*) argument selectors is large and diverse. It includes such syntactic properties as the promotion and demotion of arguments by passivization or antipassivization, the possible relativization site(s) in a relative construction, the possibility to function as either controller or controllee in various control constructions, and conjunction reduction (the interpretation of gapped arguments in coordinated clauses). See Witzlack-Makarevich 2019 for examples and further references.

2.3 *Language-internal variation in argument selection*

The generalized semantic argument roles S, A, and P, and argument selectors (for our purposes, only flagging and indexing) are not sufficient to capture language-internal variation of alignment patterns. Argument selection can vary in two primary ways: by the referential and part-of-speech properties of arguments and by various clause-level conditions.

An example of relatively straightforward variation by referential and part-of-speech information can be seen in English flagging. Some pronouns have a special P form different from the corresponding S and A form (e.g. *me* vs. *I* and *him* vs. *he*), while other pronouns have a single form for all roles (e.g. *you*, *it*). There is no such variation for nominal arguments: they never differentiate between A and P roles (e.g. *I_A kiss Lisa_P* and *Lisa_A kisses me_P*). Capturing this variation requires referencing both the person-number and the part-of-speech properties of arguments.

In addition to argument properties, a number of clausal properties are known to condition language-internal variation in argument selection. The best-known such factors are listed in (4).⁵

- (4) a. tense-aspect-mood (TAM) features
- b. the nature of the clause (main clause vs. various types of subordinate clauses)
- c. polarity
- d. scenario (co-presence of particular types of arguments in the clause)

As an example, consider the flagging of P in Aguaruna in (5) (for some generalizations, see Overall 2017). The P argument ‘chicken’ is

⁵Most of these conditions are long-established in the literature (see Dixon 1994; Bickel 2011) and have been investigated under a variety of labels, including *split alignment* (Silverstein 1976), *differential marking* (Comrie 1989), and *differential object marking* or *DOM* (Bossong 1985, 1991; Witzlack-Makarevich and Seržant 2018). The less-familiar condition is scenario (Zúñiga 2006; Witzlack-Makarevich et al. 2016), which represents a more expansive analysis of what has historically been called *hierarchical alignment* (Mallinson and Blake 1981; Nichols 1992; Siewierska 1998).

in the accusative case in (5a) and in the nominative case in (5b). This is a case of *differential object marking* (DOM). However, in contrast to the English pronouns discussed above, it is not the referential nature of the P argument that conditions the accusative case. Rather, it is exclusively the nature of the A argument that determines the marking of the P argument: if the A role references the first person singular, as in (5a), or the third person (not illustrated here), the P argument is flagged accusatively; otherwise it is flagged nominatively.

(5) Aguaruna [agua1253] (Chicham; Peru; Overall 2017, 280)

- a. atashu = n yu-a-tata-ha-i
 chicken = ACC eat-PFV-FUT-1SG-DECL
 ‘I will eat chicken.’
- b. atash yu-a-tata-hi
 chicken eat-PFV-FUT-1PL
 ‘We will eat chicken.’

In addition to the cross-linguistically recurrent conditions for variation in argument marking listed in (4), individual language descriptions occasionally include rather idiosyncratic specifications. For instance, when describing the distribution of the overt nominative flagging on S and A arguments in Achumawi [achu1247] (Palaihnihan; USA), de Angulo and Freeland (1930, p. 83) write that “subjectivity need not be indicated either, except as clearness demands it”. Such situations are recurrent and there is no principled way to compensate for gaps or vagueness in descriptive accounts.

To account for language-internal variation in argument selection, any database of alignment needs a systematic way to capture such patterns of differential argument marking. In the next section we outline the design principles of such a database, using the existing AUTOTYP alignment database (Bickel *et al.* 2022) as the starting point, and demonstrate how this design captures the multivariate nature of alignment systems.

The database presented here is not the first to collect information on alignment. WALS (Dryer and Haspelmath 2013), the first major typological database, has three features/chapters on the topic: Comrie 2013a,b with a sample of 190 languages and Siewierska 2013a with 380 languages. The more recent Grambank database (Skirgård *et al.* 2023) includes information on 2362 languages and has twelve features (GB089–GB094 and GB408–GB410) which capture similar information as WALS in a larger number of binary variables, as well as additional information about the presence of variation in marking (GB095, GB096, and GB098). Finally, Birchall 2014a, a dataset of 95 languages of South America, has a handful of alignment-related features either identical or similar to the ones in Dryer and Haspelmath 2013, as well as several related features focusing on very specific contexts (e.g. ARGEX2-7-1 asks whether verbal person marking for P is variable, obligatory or not realized when the corresponding lexical argument is present in the clause). All these databases essentially classify whole languages or whole language subsystems (e.g. pronouns in Comrie 2013b) as being of a specific alignment type selected from a previously postulated list of possible alignment types.

The database presented here took a design path quite different from the existing databases in several respects. When considered in its entirety, the phenomenon of alignment has many interacting components. We will show that it is advantageous to capture them all at once when collecting data, and to do so in such a way that multiple aggregations can be made over the same database. Our main design principles are an extension of those in AUTOTYP. We now turn to describing those principles and comparing them with those of other alignment databases.⁶

⁶The AUTOTYP database is a large-scale research program with goals in both quantitative and qualitative typology. It was launched in 1996 by Balthasar Bickel and Johanna Nichols and is thus one of the oldest typological databases still in use. AUTOTYP includes a module on grammatical relations and alignment; this has been released as Bickel *et al.* 2022. A variety of follow-up works are based on various aggregations of these data (e.g. Bickel *et al.* 2013, 2014, 2015a,b,c; Witzlack-Makarevich *et al.* 2016).

Perhaps the most common strategy in linguistic typology is to operate with variables which have a closed set of possible values. This set of possible values, either defined entirely beforehand or early on in the coding process, is essentially an etic grid which is used to categorize all individual observations. Such sets can be motivated by tradition (as in the alignment studies by Comrie 2013a,b and Siewierska 2013a), as well as by theoretical considerations or convenience. A major drawback of such pre-defined sets of possible values, especially when they are small, is that they may lack sufficient resolution to capture the full variation present in the data. For instance, the classification of a whole language as showing split alignment of indexing, as in Siewierska 2013a, does not capture what the triggers of such splits are, nor which basic alignment patterns are involved (e.g. Is it neutral and accusative? Ergative and hierarchical? etc.). This philosophy is followed by databases such as WALS (Dryer and Haspelmath 2013) and more recently by Grambank (Skirgård *et al.* 2023). AUTOTYP follows a different set of principles. Among these, the four that are most relevant for this paper are: (1) modularity, (2) autotypology, (3) late aggregation, and (4) use of exemplars (Bickel and Nichols 2002; Witzlack-Makarevich *et al.* 2022).

First, the AUTOTYP database as a whole is built in a *modular* fashion, with each module covering a typological domain. Some modules cover relatively narrow domains with just a few variables (e.g. clusivity), while others include multiple tables and several dozen variables (e.g. clause linkage). The encoding of some linguistic features may be spread across multiple modules (e.g. grammatical relations are spread among the modules on grammatical markers, predicate classes and clause linkage).

The second major design principle of AUTOTYP is *autotypology*. Autotypology means keeping variable values (and even variables themselves) flexible and open during the coding process. That is, there is no closed set of values according to which every language must be categorized. Instead, value sets and even variables can always be adjusted during coding in order to adequately capture the variability of languages. This process characterizes early stages of creation of other typological databases. This represents a radical prioritization of detailed data encoding which transparently maps to statements in reference grammars over encoding variables that match the

researcher's typological questions and previously assumed linguistic types.

The third principle is *late aggregation*. This is the principle of encoding data at a granular (autotypologized) level and only later generalizing over the data to yield cross-linguistically comparable sets of typological properties following a format familiar from conventional typological databases. Since typological categories are in principle not specified at the point of data entry, comparative typological questions are answered by querying an autotypologized database or performing data aggregations (from multiple modules if needed). As a simple example, rather than directly stating that a language has accusative flagging, the database instead lists statements about marking of various nouns and pronouns in S, A, and P roles under various conditions. The presence of accusativity can then be identified algorithmically, that is when nouns that mark S and A roles are marked differently than nouns that mark P roles. One major advantage of late aggregation is that the same data can be used to test different hypotheses and to evaluate the consequences of different operationalizations.

The fourth AUTOTYP principle is the use of *exemplars* for comparative studies, which should be extractable from the underlying data. For methodological or theoretical reasons, in some typological surveys it is desirable to have one data point per language and for these purposes one particular exemplar of a structural domain or a paradigm or a context is selected as representative for the whole domain. In other cases, a particular context or structure may be desirable as a point of comparison, without assumptions about its representativity. The use of exemplars is not unique to AUTOTYP. There are two major differences between AUTOTYP and other databases: the phase at which the exemplar comes into play; and that AUTOTYP allows for multiple exemplars during late aggregation.

The ATLAS Alignment Module largely follows the design principles of AUTOTYP outlined in Section 3, though these have been modified slightly to accommodate our coding purposes. The dataset used in this

paper is a subsample of the languages that are present in ATLAs (Inman *et al.* in prep), a global database which is focused on North and South American patterns of areality. Modifications to the AUTOTYP principles are presented in Section 4.1, an overview of the database structure is given in Section 4.2, and Section 4.3 discusses the sample and coding procedures.

Database coding

4.1

While we followed the AUTOTYP principles (Section 3) for the most part, we found it practical to depart from them in a few cases. The most significant of these departures has to do with the exhaustivity and scope: for this project, we are interested exclusively in the alignment properties of argument marking in main declarative clauses with verbal predication and with positive polarity. Thus, in a sense, one could argue that due to these limitations of scope there is some collateral violation of the principle of *autotypology*: for any contexts of the phenomenon of argument marking beyond the rather narrow predefined scope we did not expand the set of variables and their values to encode previously unencountered coding patterns. Furthermore, because our sole interest is in the alignment of morphological marking, properties of other grammatical constructions are simply not present in this database.

There are two further cases where for practical reasons we have not followed the principle of autotypology.

First, it is impossible to know in advance all possible variables by which morphological alignment might vary in a sample. The most typical conditions are properties such as TAM and predicate class, and, following the autotypology principle, we have left the possible values of these variables open-ended during coding. However, there are many other possible sites of variation (e.g. word order, the presence or absence of an overt NP, or unknown or insufficiently described conditions). To track these conditions on alignment variation, we have created a single variable called “Miscellaneous conditions” which is used to cover all of these “other” conditions. The set of values that “Miscellaneous conditions” can accommodate is open ended and should in principle be split into separate variables following the

principle of autotypology. However, we have kept this as a single variable since these various conditions are not the primary target of this study.

Second, we included a convenience variable⁷ explicitly indicating a highly specific exemplar of flagging and indexing chosen beforehand, instead of computing it after the fact. This adds to rather than detracts from the AUTOTYP way of dealing with exemplars outlined in Section 3, since it only abstracts in a non-algorithmic fashion over information that is already present in other variables. The exemplar we chose in this project is defined by Birchall (2014b, 24–25) (following Lazard 2002, 252). We have adopted and expanded on Birchall’s definition and termed it the *exemplar declarative main clause*. This exemplar has the following properties:⁸

- (6) Exemplar declarative main clause:
 - a. The clause represents a real event (not prospective, not imagined) and is declarative.
 - b. The clause is not embedded or a complement of another clause.
 - c. The event described in the clause is discrete, perfective or completive, and not ongoing or incomplete.
 - d. The clause has positive polarity and is not negated.

Since morphosyntactic alignment is a phenomenon that can vary depending on the characteristics of the arguments, in addition to defining the exemplar clause, the exemplar S, A, and P roles are defined in (7).

⁷ By “convenience variable” we mean a variable that is not strictly necessary and does not encode any additional information. As we will outline below, the exemplar variable could in principle be derived algorithmically from the other variables present in the database, although such an algorithm would be cumbersome.

⁸ There are consequences to adopting any exemplar. In our case, the definitions in (6) and (7) will preferentially select for accusative alignments, as many languages with split-S marking mark S arguments if they control events the same as A arguments, and thus all these languages will be considered as showing accusative alignment in the exemplar case. We have captured the existence of such systems by making sure to encode monovalent predicate classes where the S lacks control (see the discussion on `Predicate_class` in Appendix A.2).

- (7) Exemplar S, A, and P arguments:
- a. The S argument is a human that voluntarily performs and controls the event.
 - b. The A argument is a human that voluntarily performs and controls the event.
 - c. The P argument is well-individuated, human, and is actually affected by the event.

It is in principle possible to algorithmically derive this exemplar from the TAM, predicate class, and miscellaneous conditions defined for each context. However, because the possible values of these variables are all open-ended, the relevant algorithm would need to include a constantly updated classification of all these conditions (and possibly their interactions) to allow the extraction of only those contexts which represent the exemplar case. Encoding the exemplar in a convenience variable avoids the need to create and continuously update such a list. Though we have encoded this exemplar variable according to the properties defined in Birchall 2014b, this kind of information could be encoded for other exemplars, with each exemplar encoded in a separate convenience variable.

A practical decision was needed as to how and whether to encode the absence of overt marking (or “zero marking”). For nouns and pronouns, we coded contexts for each role S, A, and P, whether they had overt flagging or not. All zero marking in flagging, therefore, is coded explicitly. However, we determined that it was not feasible to do this for indexing. If a language has several slots for indexing, e.g. different slots on the verb for different persons and roles, then there could be many zeros simply indicating that a particular person is absent from a context. In more complex cases, it is unfeasible to code all zeros, or doing so would require making decisions about possibly indeterminate properties (for example, how many slots are present in a certain configuration). There is also a theoretical decision to be made, about whether there is a “true zero” which means something, or if marking is simply absent. This cannot always be determined from available sources.

For the coding of zeros in indexing, we adopted the policy that they need not be explicitly coded, but could be. However, there are some cases where the coding has to be explicit, with a phonologically

zero selector: (1) when a zero is the only reflex of a particular referential type (e.g. 3rd person singular is not marked), or (2) when a zero marker contrasts with an overt marker under certain conditions (e.g. a 3rd person index which is phonologically overt under some conditions and zero under others).⁹ However, in other cases, such as the exceptionless absence of indexing for the P role, or the absence of marking in a particular slot in a particular scenario, we allowed for this to be coded explicitly or not, depending on the ease and preference of the coder. This creates a certain level of inconsistency in our database: Sometimes these zeros (both the lack of indexing for a role and the lack of overt marking in a particular case) are present, and sometimes they are not. But in terms of database interpretability, nothing is lost: The absence of explicit information about the indexing of S, A, or P arguments means that there is no overt marking.

4.2

Database structure

The ATLAS Alignment Module conforms to the CLDF standard (Forkel *et al.* 2018) and is composed of three basic csv files (`contexts.csv`, `selectors.csv`, and `languages.csv`) and the `metadata.json` file that describes how the csv files are interrelated. As the CLDF format is customizable and extendable, further information can be added in the form of new columns and even new tables.¹⁰ As Section 5 shows, we add such derived columns and tables as we proceed with querying the database to create data aggregations at different levels (for an overview of the database structure, see Figure 1).

Each of the basic csv files are briefly described below in Sections 4.2.1–4.2.3, with an overview of the most important columns

⁹This means that the full list of referential types indexed in a language is always available in `contexts.csv`, unless they behave uniformly in terms of alignment (see Section 4.2.1). In order to perform meaningful aggregations on complex indexing systems (see Section 5), we need a record of all referential types the indexing systems distinguish no matter whether they are overtly marked or not. Thus each referential type must have at least one context indicating its existence. The other possibility would be to have a separate table listing all referential types for all languages.

¹⁰In the remainder of the paper we use monospace typeface for file names and column headers and we enclose variable values in `<angle brackets>`.

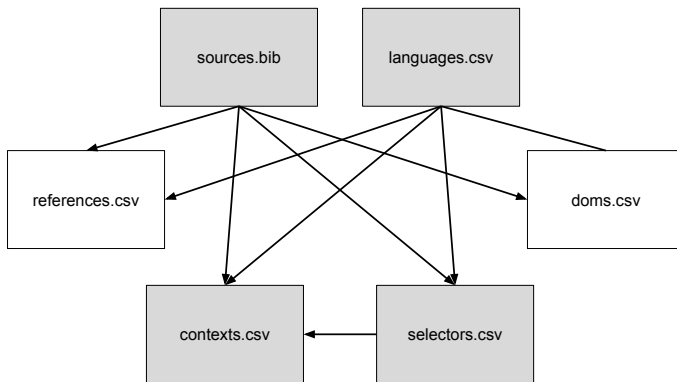


Figure 1:
Representation of CLDF
database. Basic files
with raw data are
represented by gray
rectangles, additional files
populated with scripts by
white rectangles. Lines
show one-to-one
relationships, and arrows
one-to-many

and coding decisions for each, along with excerpts from each csv file that present the corresponding content for two illustrative languages: Bilua [bilu1245] (isolate; Papua New Guinea and Solomon Islands) and Awa-Cuaiquer [awac1239] (Barbacoan; Colombia and Ecuador). Bilua is straightforward as far as alignment is concerned: there is no flagging for nouns or pronouns, and S and A roles are indexed with a paradigm of proclitics and P with a paradigm of enclitics (see example (8)). The last three rows in Table 1 represent indexing in Bilua: the same selector <bilua1245-s-a-proclitics-indexing-marker> (corresponding to the proclitic paradigm) is used for both S and A roles and appears in the <clitic -1> slot. The enclitic paradigm for the P role is a separate selector <bilua1245-p-enclitics-indexing-marker> and appears in the <clitic 1> slot.

- (8) Bilua [bilu1245] (isolate; Papua New Guinea and Solomon Islands; Obata 2003, 309)¹¹

ko = rere = a inio ko = pa zuzue = v = a
3SG.F = run = PRS SEQ 3SG.F = PROS hug = 3SG.M.O = PRS

‘She ran and then she hugged him.’

¹¹Special glosses for Bilua which extend the Leipzig Glossing Rules (Comrie *et al.* 2008) are: SEQ: sequential coordinator, PROS: prospective marker. The present tense marker in this example is used as historical present tense, and is thus translated using the past tense.

Awa-Cuaiquer on the other hand is more complicated: it has co-argument sensitivity as well as both a split-S system and a fluid-S system, where fluidity applies only to S arguments of stative verbs and only matters for the markers of the 1st person. The example (9) below is represented with three different contexts in Table 1:

- in the line with ID <awac1239-5>, the reference is a high noun (humans in Awa-Cuaiquer) in the P role, which is marked with the accusative case, irrespective of the A coargument being a noun or a pronoun;
- in the line with ID <awac1239-8>, the reference is a pronoun in the A role, which is unmarked, irrespective of the P coargument being a noun or a pronoun;
- in the line with ID <awac1239-15>, the reference is a non-locutor (in example 9, a second person) in the A role with another non-locutor¹² (in the example, a third person) in the P role. The A argument is indexed on the verb with the suffix *-zi*, which is specific to contexts where no locutor (first person) is involved.

(9) Awa-Cuaiquer [awac1239] (Barbacoan; Colombia and Ecuador; Curnow 1997, 199)

nu = na Juan = ta pyan-ti-zi
2SG.(NOM) = TOP Juan = ACC hit-PST-NONLOCUT

‘You hit Juan.’

For a more detailed description and explanation of all the values for each column, see the Appendices.

4.2.1

contexts.csv

In the contexts.csv table (see Table 1),¹³ each row represents a context involving either one argument (S in the case of monovalent verbs) or two arguments (A and P in the case of bivalent verbs), and exactly

¹²In Awa-Cuaiquer, indexing distinguishes only 1st person (locutor) from 2nd/3rd person (non-locutor).

¹³In Tables 1–3, Table 5, and Table 7, some columns have been omitted for readability.

Table 1: Excerpt from contexts.csv

ID	Selector_ID	Slot	Role	Reference	Co-argument ₁	Co-argument ₂	Exemplar	Predicate _{class}	Miscellaneous _{condition}
awac1239-1	awac1239-no-flagging		S	Noun-high	NA		any	default	
awac1239-2	awac1239-no-flagging		A	Noun-high	P		any	default	
awac1239-3	awac1239-no-flagging		S	Noun-low	NA		any	default	
awac1239-4	awac1239-no-flagging		A	Noun-low	P		any	default	
awac1239-5	awac1239-acc-marking-flagging		P	Noun-high	A		any	default	
awac1239-6	awac1239-no-flagging		P	Noun-low	A		any	default	
awac1239-7	awac1239-no-flagging		S	Pro	NA		any	default	
awac1239-8	awac1239-no-flagging		A	Pro	P		any	default	
awac1239-9	awac1239-acc-marking-flagging		P	Pro	A		any	default	
awac1239-10	awac1239-suffix-s-1p-indexing-marker	1	P	1	A		any	default	any
awac1239-11	awac1239-suffix-s-1p-indexing-marker	1	S	1	NA		non-exemplar	stative	unknown condition 1
awac1239-12	awac1239-suffix-w-1s-a-indexing-marker	1	A	1	P		any	default	any
awac1239-13	awac1239-suffix-w-1s-a-indexing-marker	1	S	1	NA		any	default	control
awac1239-14	awac1239-suffix-w-1s-a-indexing-marker	1	S	1	NA		non-exemplar	stative	unknown condition 2
awac1239-15	awac1239-suffix-zi-2-3s-a-indexing-marker	1	A	2/3	P		any	default	any
awac1239-16	awac1239-suffix-zi-2-3s-a-indexing-marker	1	S	2/3	NA		any	default	control
awac1239-17	awac1239-suffix-zi-2-3s-a-indexing-marker	1	S	2/3	NA		non-exemplar	stative	any
bilu1245-1	bilu1245-no-flagging		S	any	NA		any	default	
bilu1245-2	bilu1245-no-flagging		A	any	P		any	default	
bilu1245-3	bilu1245-no-flagging		P	any	A		any	default	
bilu1245-4	bilu1245-p-enclitics-indexing-marker	clitic 1	P	any	A		any	default	
bilu1245-5	bilu1245-s-a-proclitics-indexing-marker	clitic -1	S	any	NA		any	default	
bilu1245-6	bilu1245-s-a-proclitics-indexing-marker	clitic -1	A	any	P		any	default	

one selector which is associated with this particular context. Each argument present in a particular context is referred to in terms of its role (see Section 2.1). The argument selector (a morpheme or a paradigm of morphemes) associated with a context is identified by a selector ID, which is linked to the `selectors.csv` table, where selector-specific information is collected. Contexts are language-specific, and the language that a context belongs to is specified through a language ID (note that this column has been omitted in Table 1, but the Glottocode is still visible in the ID column).

Because all contexts are associated with exactly one selector, they must minimally be specified for the argument roles and references involved. However, a context may require more information (such as slot, TAM or predicate class) to distinguish it from other contexts in the language which are associated with different selectors.¹⁴

In most languages, morphological slot can be seen as a property of the selector in question, but this is not always the case. In some languages, such as Puinave [puin1248] (isolate; Colombia and Venezuela), the same paradigm of person indices is used for both the A and P roles but appears in different slots on the verb (Girón Higueta 2008). To have a unified approach, we treat the slot as a property of the context and the same selector can appear in different slots depending on the context.

Another case where more information is needed to identify a context is when a language uses different verbal paradigms for indexing person-number values in different tenses, as is the case for many Indo-European languages. These different paradigms correspond to different selectors, and so the context must be able to distinguish when one paradigm or the other is used. This is accommodated by the dedicated column for TAM. Separate columns for predicate class, co-arguments, and miscellaneous conditions accommodate other cases where contexts may differ. This structure proved sufficient to capture marking variation in the languages we have encountered.

For practical reasons, we do not differentiate between contexts when there is no difference in terms of alignment. For example, we do not list all person and number combinations for person subject indexes

¹⁴Note that in Table 1, the TAM column has been omitted because it was not relevant for the languages exemplified.

in a language such as Bilua, where there are two paradigms of clitics that behave uniformly (Obata 2003, 49, 303, 309). In such cases, each row represents a bundle of contexts that have in common the same argument role (see Table 1). Thus, in the Bilua example, there are three rows in `contexts.csv`: two that correspond to the subject clitic paradigm (one for indexing S and one for indexing A) and one row for the object clitic paradigm (P indexing). The roles themselves may be combined in one context row in cases of complete absence of verbal indexing for any role.¹⁵

As a final note, (person indexing) selectors which function in certain contexts as portmanteaus (i.e. they index both A and P arguments)¹⁶ have two entries in the `contexts.csv` table. Since an entry in the `contexts.csv` table represents the marking of both a role and a referential type, such selectors have two entries for the same scenario: one for marking the A role given the appropriate P as its co-argument and another one for marking the P role given the appropriate A as its co-argument. Though this may seem like a kind of double-coding, it is analogous to a single selector used to mark both S and A roles.

`selectors.csv`

4.2.2

In the `selectors.csv` table, each row corresponds to a morpheme or a paradigm of morphemes (see Table 2). The label of this morpheme or paradigm is given in free form as its `Selector_label`,

¹⁵ We only allow for this collapsing of argument roles in the case of an absence of indexing, and not in the case of an absence of flagging. Unlike verbal indexing, which can be completely absent in a language, flagging is almost always present if we take into account all argument roles. It is very common that other argument roles currently not coded in our database, such as G (goal) or T (theme), have distinct flagging, even if the S, A, and P argument roles do not.

¹⁶ The property of a selector behaving as a portmanteau is commonly seen as inherent to the selector, e.g. an indexing marker is either a simple or a portmanteau marker. However, in some languages the same selector may function as a simple marker in some contexts and as a portmanteau marker in others. As an example, in Huastec [huas1242] (Mayan; Mexico), the marker *tu=* indexes the 1st person plural P role. However, it is also used in all cases where 1st person A acts on 2nd person P (Edmonson 1988, pp. 114–115). In the former case, the morpheme behaves as a simple P marker; but in the latter case, it can only be understood as a portmanteau. We have therefore opted for considering portmanteau behavior as a property of the context rather than the selector.

Table 2: Excerpt from `selectors.csv` corresponding to the contexts given in Table 1

Glottocode	Selector_type	Selector_label	Marker_type	Features
awac1239	flagging	ACC marking	overt	
awac1239	flagging	NO_FLAGGING	zero	
awac1239	indexing marker	suffix -s 1P	overt	person
awac1239	indexing marker	suffix -w 1S/A	overt	person
awac1239	indexing marker	suffix -zi 2/3S/A	overt	person
bilu1245	flagging	NO_FLAGGING	zero	
bilu1245	indexing marker	P enclitics	overt	person + number
bilu1245	indexing marker	S/A proclitics	overt	person + number

which could be an abstract value (like `<ergative suffix>`) or a more concrete one (such as the phonological shape of a person indexing morpheme, e.g. `<mü- 3sgA>`). Each selector is given a value for its `Selector_type` which specifies whether this selector is used for flagging or indexing, and a `Marker_type` which specifies if it is phonologically `<overt>` or `<zero>`. The `Selector_type` can be `<flagging>`, `<indexing marker>`, or `<indexing trigger>`, the latter of which is a special type indicating a lack of indexing for a role (and thus always has `Marker_type <zero>`). Zero morphemes that encode a specific referential type have `Selector_type <indexing marker>` or `<flagging>` and `Marker_type <zero>`, while zeros that represent the lack of indexing in general, or the lack of indexing for a particular role, are always `Selector_type <indexing trigger>`. A consistent selector label `<NO_FLAGGING>` is used for the absence of flagging of a specific argument role.

The `selectors.csv` table includes other information about selectors, such as what features they index (e.g. number, person). Selectors are linked to the language they belong to by the `Glottocode` column.

4.2.3

languages.csv

In the `languages.csv` table, each row is a language characterized by a unique ID and associated information such as family membership, geographical coordinates etc. These data are following Glottolog 4.8

Table 3: Excerpt from `languages.csv`

Glottocode	Name	Macroarea	Latitude	Longitude	Family
awac1239	Awa-Cuaiquer	South America	1.21652	−78.3401	Barbacoan
bilu1245	Bilua	Papunesia	−7.92388	156.663	

(Hammarström *et al.* 2023). There is also a comment field for any unstructured information on the language as a whole, such as the presence or absence of co-referential personal pronouns. An excerpt from the `languages.csv` table is given in Table 3.

Sample and data collection

4.3

We have selected a geographically-balanced diversity sample of 84 phylogenetically unrelated languages (according to Glottolog 4.8, Hammarström *et al.* 2023), equally distributed among each of the world’s six macroareas (Hammarström and Donohue 2014). All of our figures and results are based on this sample of languages, the full list of which can be found in the Supplementary Materials in the `languages.csv` file.

The data collected for this dataset were extracted from primary source documents, mostly from reference grammars and linguistic articles. Only occasionally did we consult native speakers and language specialists (via personal communication).

During data collection, in addition to the entries in our database structure, we created a more human-readable summary of each language’s flagging and indexing patterns, complete with detailed references and quotes. This summary was used in team discussions, as well as a reference point for necessary adjustments during autotypologizing. Data consistency during the coding procedure was aided by custom scripts, which reported on definitionally impossible entries (e.g. a claim that two morphemes occupy the same slot on the verb in the same context, or that a noun was marked with two different cases in the same context), which were then corrected manually.

5 DATABASE QUERYING AND RESULTS

The database structure described in Section 4.2 does not itself answer any specific typological questions, but the database can be queried to answer a large variety of possible questions. We exemplify a few of the most typical ways of calculating alignment statements. Some of these queries match familiar alignment statements present in other databases, whereas others are impossible to retrieve from statements in other databases. The examples below are by no means exhaustive of the typological properties that can be extracted from our database.

We organize these alignment properties into different levels of linguistic structure. It is possible to specify typological questions about alignment at the level of the language (Section 5.1), at the level of individual argument selectors (Section 5.2), at the level of individual referential types (Section 5.3), and at the level of argument roles (Section 5.4). All queries and aggregations are implemented in individual functions in the accompanying `alignment_aggregation.py` file in the Supplementary Materials.

5.1 *Language-level aggregation*

Several properties of alignment can be established at a language-wide level, without having to calculate per-selector, per-referent, or per-role information. We have defined queries for five of these and implemented them in the `basic_language_level` function in `alignment_aggregation.py`:

- (10) a. the presence of flagging for core arguments
- b. the presence of indexing
- c. the features which are targeted by indexing, if there is any
- d. the presence of an alignment split conditioned by TAM properties
- e. the presence of a split-S system

The presence of overt argument flagging (10a) is retrieved from the `selectors.csv` table by querying, for each language, whether

there are any selectors for which the `Selector_type` is `<flagging>` and the `Marker_type` is `<overt>`. The presence of indexing (10b) is likewise retrieved from the `selectors.csv` table by querying for rows in which `Selector_type` is `<indexing marker>` and `Marker_type` is `<overt>`. The features targeted by indexing (10c) are retrieved by concatenating all unique non-`<NA>` values in the `Features` column for all indexing selectors of this language.¹⁷

The presence of an alignment split conditioned by TAM properties (10d) is retrieved from the `contexts.csv` table by querying, for each language, whether there is more than one value present in the `TAM` column. The presence of multiple values indicates that TAM properties are relevant for an alignment split.

Finally, the presence of a split-S system (10e) is also retrieved from the `contexts.csv` table by querying for rows marking the `S` role which have a `Predicate_Class` value other than `default`.

Some of these properties, such as the presence of a split-S system, occur frequently in studies on alignment, while others, such as the features targeted by indexing, do not. However, answers to both typical and less typical questions can be extracted easily from our database. We can additionally address typological properties at other levels of organization, below the level of the language as a whole, as we will see in the next sections.

The results of these queries for each language are written to `structure-cldf/values.csv`, in accordance with the CLDF format, and another version is optionally written to the non-CLDF compliant `human-readable.csv`, which is organized by language rather than by language and parameter. Statistics can then be calculated on this output.¹⁸ Although the sample size for this dataset is relatively small,

¹⁷ While it is possible to aggregate these values (e.g. a language with a selector which targets `<number>` and another selector which targets `<person>` could be aggregated into `<person+number>`), we chose to keep them separate (e.g. such a language would have a value `<person;number>` for this query).

¹⁸ All statistics are implemented in the `write_summary_statistics` function of `alignment_aggregations.py`, which reads the CLDF-compliant csv output of each level of aggregation and calculates summary statistics. These statistics are written to file at `summary.csv`, which can be accessed in the Supplementary Materials.

some summary statistics of these language-level aggregations are presented in Table 4.

Table 4: Selected language-level results (N = 84)

Property	Count	Frequency
Presence of argument flagging	47	56%
Presence of argument indexing	59	70%
TAM-based alignment split	3	4%
Split-S system	9	11%
Person + number always indexed together (if indexing present)	42	71%

5.2

Selector-level aggregation

In addition to alignment properties at the language level, it is possible to derive alignment statements at the level of individual argument selectors. Selectors mark roles (S, A, or P), either as argument flagging (on the NP) or indexing (on the verb/clause), and an individual selector may mark multiple referential types (e.g. the same verbal index might be used for both 3rd person singular and 3rd person plural A arguments).

The first question that can be answered about an argument selector is: “Which role(s) are marked by this specific marker?” For example, an argument selector may mark S and A roles, but not P; or S and P, but not A. Once it has been determined which argument roles a selector marks, an alignment statement can be calculated for that selector. This selector-based “alignment” is not quite the same as reference-based alignment, which is what is prototypically referred to by the term (see Sections 2.1 and 5.3). At the level of an individual selector (disregarding for the time being what it is referencing), it either marks a particular role, or it does not (e.g. a specific case suffix either marks S arguments or it does not). There is in this sense no such thing as a tripartite alignment for selectors: since its presence is a binary value, it is impossible to have the state $S \neq A \neq P$. For the same reason, selectors which function exclusively as portmanteaus (such as a morpheme marking 2>1) will always have a horizontal alignment

at the selector level (the marking of A and P but not S). This differs from reference-based alignment, where a horizontal alignment means that for a given reference (e.g. 2sg) the A and P (but not S) roles are marked by the same morpheme.

Note that zero selectors (the absence of marking) can also have an alignment. A zero-marked nominative case (contrasting with an overtly-marked accusative case) still has a selector-based (nominative-)accusative alignment. The only case in which a non-overt selector does not have an alignment (Alignment is <NA>) is when a role is not marked at all. As we discussed in Section 4.1, this is possible with indexing (the selector type <indexing trigger>), but not with flagging.¹⁹ Selector-based alignment is closely related to the concept of trigger potential (Siewierska 2003; Bickel *et al.* 2013), because it describes which roles *can* trigger the appearance of a particular morpheme. As such, selector-based alignment can only have the values neutral, accusative, ergative, and horizontal.

For the selector-level aggregation, we wrote queries to add four columns to the `selectors.csv` table (see Table 5). The first three columns, `S_references`, `A_references`, and `P_references`, keep track of which references a selector marks. The values of these columns are generated by looking in the `contexts.csv` table for all instances of a given `Selector_label`, and entering into the appropriate column in the `selectors.csv` table which referential types that selector can reference. If a referential type is conditioned by a co-argument, they are concatenated, e.g. if a selector only marks 1st person A when P is 2nd person, the value entered is <1_2>. If no reference is marked for that role, the value <NONE> is entered in the references list.

Finally, the fourth column, `Alignment`, is added. The value of this column is calculated based on the presence or absence of referential types in the `S_references`, `A_references`, and `P_references` columns, regardless of what values are present. For example, if a

¹⁹ As explained in Section 4.2.1, we consider slot a property of the context, rather than of the selector. Therefore, for each language there is at most one zero selector for flagging and one for indexing and they can appear in different slots. These zero selectors are of the type <flagging> or <indexing marker> respectively and are treated identically to other selectors of the same type.

Table 5: Excerpt from selectors.csv with added columns from the selector-level queries

Glottocode	Selector_type	Selector_label	S_references	A_references	P_references	Alignment
awac1239	flagging	ACC marking	NONE	NONE	Noun-high; Pro	accusative
awac1239	flagging	NO_FLAGGING	Noun-high; Noun-low; Pro	Noun-high; Noun-low; Pro	Noun-low	neutral
awac1239	indexing marker	suffix -s 1P	1	NONE	1_2/3	ergative
awac1239	indexing marker	suffix -w 1S/A	1	1_2/3	NONE	accusative
awac1239	indexing marker	suffix -zi 2/3S/A	2/3	2/3_2/3	NONE	accusative
bilu1245	flagging	NO_FLAGGING	any	any	any	neutral
bilu1245	indexing marker	P enclitics	NONE	NONE	any	accusative
bilu1245	indexing marker	S/A proclitics	any	any	NONE	accusative

Variable	Value	Count	Frequency
Selector flagging of S	True	16	26%
Selector flagging of A	True	30	49%
Selector flagging of P	True	35	57%
Selector flagging alignment	Accusative	40	66%
	Ergative	17	28%
	Neutral	3	5%
	Horizontal	1	2%
Selector indexing of S	True	222	58%
Selector indexing of A	True	247	65%
Selector indexing of P	True	204	54%
Selector indexing alignment	Accusative	238	63%
	Neutral	53	14%
	Ergative	45	12%
	Horizontal	44	12%

Table 6:
Aggregations
of flagging
selectors (N = 61)
and indexing
selectors
(N = 380)

particular selector has a non-`<None>` entry in the `S_references` and `A_references` columns, but `<None>` in the `P_references` column, then its value for `Alignment` is `<accusative>`, even if the values present in the `S_references` and `A_references` columns are different.

As we did with language-level aggregation, we present summary statistics at the level of selectors. Here, we have only calculated these statistics for overt markers. These statistics could also be balanced per language, so that languages with many selectors are weighted evenly with languages that have fewer. We present the unbalanced, selector-level statistics for some of these properties in Table 6.

Reference-level aggregation

5.3

Another possible level of aggregation is at the level of referential types. For pronouns and verbal indexing, the relevant referential types are the various person-number combinations attested in the language, while the relevant referential types for nouns are the different groups of nouns (if any) that behave uniformly as far as argument flagging is

concerned (e.g. masculine, feminine, singular, plural, etc.). Thus, it is possible for a language to have e.g. a tripartite alignment for first person singular indexing (different selectors are used for each of the S, A, and P roles), but an accusative alignment for second person singular indexing (S and A roles are indexed with the same selector, while P has a distinct one). Similarly, nouns in the singular may exhibit accusative flagging (nominal S and A arguments are in the nominative case, whereas nominal P arguments are in the accusative case), while nouns in plural may have neutral flagging (the same nominative form is used for all three roles). The reference-level alignment can also be different under different conditions, such as TAM or different predicate classes. In such cases a reference-level alignment is calculated for each of those different conditions.

The reference-level aggregation is implemented in the `reference_alignment.py` script, available in the Supplementary Materials. This script extracts, per language, how each combination of role and referential type is marked. If further conditions are relevant, such as TAM, then the marking of each role and reference combination is calculated per condition. In cases of co-argument sensitive marking, there is no single marking strategy for a role and reference combination, but several, dependent on the co-argument. In such cases, the script extracts a series of marking strategies depending on the co-argument. A detailed example of the script functionality and code flow can be found in Appendix C.

The results of the aggregation at the reference level are written to a separate `references.csv` file (see Table 7). Each row in this table represents a particular referential type of a particular language under specific conditions. Each row is identified with a unique ID and is linked with the corresponding language through the `Glottocode` column, while the relevant referential type is listed in the `Referential_type` column. The `references.csv` table also includes several additional columns that specify the conditions (`Monovalent_predicate_class`, `Bivalent_predicate_class`, `TAM`, `Condition`), one column per role (S, A, and P), and two alignment columns (`Alignment` and `Alignment_not_local`), which are calculated based on the role columns. For a more complete description of the `references.csv` table, its columns and possible values, see the Appendices.

Table 7: Excerpt from references.csv

Glottocode	Domain	Referential_type	Exemplar	Movavalent_predicate_class	Condition	S	A	P	Alignment	Alignment_not_local
awac1239	Verb	1	exemplar	default	control	suffix -w 1S/A_overt	suffix -w 1S/A_overt	suffix -s 1P_overt	accusative	accusative
awac1239	Verb	2/3	exemplar	default	control	suffix 2/3S/A_overt	suffix -zi 2/3S/A_overt_coarg:2/3 ; INFERRED_NULL_zero_coarg:1	INFERRED_NULL_zero	sensitive	sensitive
awac1239	Verb	1	all	stative	unknown condition 1	suffix -s 1P_overt	suffix -w 1S/A_overt	suffix -s 1P_overt	ergative	ergative
awac1239	Verb	2/3	all	stative	unknown condition 1	suffix 2/3S/A_overt	suffix -zi 2/3S/A_overt_coarg:2/3 ; INFERRED_NULL_zero_coarg:1	INFERRED_NULL_zero_coarg	sensitive	sensitive
awac1239	Verb	1	all	stative	unknown condition 2	suffix -w 1S/A_overt	suffix -w 1S/A_overt	suffix -s 1P_overt	accusative	accusative
awac1239	Verb	2/3	all	stative	unknown condition 2	suffix 2/3S/A_overt	suffix -zi 2/3S/A_overt_coarg:2/3 ; INFERRED_NULL_zero_coarg:1	INFERRED_NULL_zero	sensitive	sensitive
awac1239	Verb	1	all	default	control	suffix -w 1S/A_overt	suffix -w 1S/A_overt	suffix -s 1P_overt	accusative	accusative
awac1239	Verb	2/3	all	default	control	suffix 2/3S/A_overt	suffix -zi 2/3S/A_overt_coarg:2/3 ; INFERRED_NULL_zero_coarg:1	INFERRED_NULL_zero	sensitive	sensitive
awac1239	Noun	Noun-high	exemplar	default		NO_FLAGGING_zero	NO_FLAGGING_zero	ACC marking overt	accusative	accusative
awac1239	Noun	Noun-low	exemplar	default		NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
awac1239	Noun	Noun-high	all	default		NO_FLAGGING_zero	NO_FLAGGING_zero	ACC marking overt	accusative	accusative
awac1239	Noun	Noun-low	all	default		NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
awac1239	Pro	Pro	exemplar	default		NO_FLAGGING_zero	NO_FLAGGING_zero	ACC marking overt	accusative	accusative
awac1239	Pro	Pro	all	default		NO_FLAGGING_zero	NO_FLAGGING_zero	ACC marking overt	accusative	accusative
bilu1245	Verb	any	exemplar	default		S/A proclitics_overt	S/A proclitics_overt	P enclitics overt	accusative	accusative
bilu1245	Verb	any	all	default		S/A proclitics_overt	S/A proclitics_overt	P enclitics overt	accusative	accusative
bilu1245	Noun	any	exemplar	default		NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
bilu1245	Noun	any	all	default		NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
bilu1245	Pro	any	exemplar	default		NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
bilu1245	Pro	any	all	default		NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking

The Alignment column takes into account all contexts, while the Alignment_not_local column excludes from the calculation local scenarios (1st person acting on 2nd and vice versa), since in many languages politeness may have affected the alignments of such scenarios (see e.g. Heath 1998; DeLancey 2021). For example, if the S and A column for the referential type <noun> are both <NOM_zero> and the P column is <ACC_overt>, then the Alignment would be <accusative>, as would be the Alignment_not_local. If there is co-argument sensitivity for any role, then the Alignment is given the value <sensitive>.²⁰ If this sensitivity is due only to local scenarios, then the Alignment_not_local column would have a non-sensitive value.²¹ Finally, if all markers involved in the flagging of a particular referent are non-overt, the resulting Alignment is <no marking>, a special type of neutral alignment. Another case of neutral alignment but with overt markers is attested more often in indexing than flagging, namely, in cases where the same set of markers is used for all roles. We call this alignment pattern <overt neutral>.

With the references.csv table, we can once again perform summary statistics, in this case over all referential types. First, we can calculate, per language, what the most common reference-based alignment pattern is (using the Alignment column rather than the Alignment_not_local one), weighting all referential types equally. The results are summarized in Table 8, which reports for each selector subtype (Flagging on nouns, Flagging on pronouns, Indexing) all the most frequent reference-based alignments per language that occur at least 5% of the time in our sample.

Another kind of aggregation that can be done at the reference level (and without further aggregation per language) is the presence of paradigmatic zeros in indexing, i.e. which referential types are not

²⁰ The value <sensitive> is not a proper alignment the way that accusative, ergative, etc. are. It is a bundle of different alignments that are dependent on co-argument references. It is in principle possible to further decompose this state into individual alignment statements not per reference but per combination of reference and co-argument reference, as in Witzlack-Makarevich *et al.* 2016.

²¹ In Awa-Cuaiquer it is not possible to assign a scenario involving a 1st person as mixed or local, since second and third persons are not distinguished. We have opted conservatively to keep all potentially non-local scenarios, and the <sensitive> alignment is retained; see Table 7.

Variable	Value	Count	Frequency
Flagging on nouns	no marking	44	52%
	accusative	14	17%
	ergative	9	11%
	no marking/accusative	7	8%
Flagging on pronouns	no marking	41	49%
	accusative	27	32%
	ergative	8	10%
Indexing	accusative	33	39%
	no marking	27	32%
	co-argument sensitive	14	17%

Table 8:
Most frequent
reference-based
alignments per
language (> 5%)

Person	Role	Count	Frequency
Zero indexing for 1	S	6	3%
	A	7	4%
	P	14	9%
Zero indexing for 2	S	0	0%
	A	17	10%
	P	17	13%
Zero indexing for 3	S	55	22%
	A	65	26%
	P	51	23%

Table 9:
Zeros in indexing
by person reference

indexed for S, A, and P roles, while other referential types are marked under the same conditions. Table 9 presents cases of zero indexing broken down by person (without distinguishing number, i.e. 2sg and 2pl each count as independent examples of indexing of 2). As Table 9 shows, in our sample the P role more frequently lacks indexing than S and A roles, as is expected from previous research (Siewierska 2013b). Furthermore, the 3rd person more frequently lacks indexing than 1st and 2nd (see e.g. Bickel *et al.* 2015c).

5.4

Role-level aggregation

Information about argument marking can also be aggregated at the level of the role (S, A, P). Such aggregations are not alignment, as typically conceived, since they concern exclusively the manner of marking of individual argument roles, i.e. the manner of S marking (on its own), the manner of A marking, and the manner of P marking. This aggregation allows one to capture the various patterns of differential argument marking (see Witzlack-Makarevich and Seržant 2018 for a recent overview). The best studied type of differential argument marking which corresponds to this level of aggregation is differential object marking (or DOM, see Bossong 1985, 1991). In addition to DOM, there are differential A marking (or DAM) and differential S marking (including split-S or active-stative systems, see Section 5.1).

For this paper, we have only aggregated information about DOM. For our present purposes, we are considering DOM “in the broad sense” (see Witzlack-Makarevich and Seržant 2018), that is, we treat as DOM any case of variation in the marking of the P argument irrespective of the condition triggering it, such as different referential types (e.g. definite vs. indefinite), different TAM of the clause, etc, so long as this change is also accompanied by a change in alignment.²² It is possible for a language to have complex systems of DOM with more than one factor conditioning the split, e.g. person and TAM. In these cases, we present the combined conditioning factors causing the split.

The presence of DOM is extracted from the `contexts.csv`, `selectors.csv`, and `references.csv` tables. First, we select, per language, all the rows in the `contexts.csv` table which have their Role marked as `<P>`, and which encode flagging information (the associated selector in the `selectors.csv` table has the `Selector_type` `<flagging>`). If these rows contain different selectors and at least one has `Marker_type` `<overt>` (i.e. not all are `<zero>`), then the `references.csv` table is checked for whether these P selectors are as-

²² We consider `<overt neutral>` and `<no marking>` as the same, since they are both subtypes of neutral alignment.

sociated with different alignments. If this is the case, then a language with DOM has been found.²³

Once a language is established as having DOM, then the conditions which cause the differential P marking are calculated. The potential set of DOM-triggering conditions in the `contexts.csv` table can be found in the columns `TAM`, `Reference`, `Co-argument_reference`, `Miscellaneous_condition`, and `Predicate_class`. If one of these columns has values which are each associated with unique P markers and different alignment statements, then that column is the conditioning environment for the DOM. However, as we mentioned above, it is possible that the DOM is conditioned by two (or even more) conditions; if single columns fail to distinguish between different P markings, then each possible combination is tested. The full details of this extraction are given in the `alignment_aggregation.py` script.

Once calculated, the different DOMs are output to `doms.csv`. Each row represents a single DOM and indicates the language (Glottocode) and the conditioning factor which causes it (the `Conditioning` column), e.g. a different reference, TAM, and so forth. In addition, there is a series of columns for each marking, the set of alignments it is associated with, and the corresponding conditions (see the Appendices for more details).

Table 10 shows a simplified example taken from the table generated by our DOM aggregation. Central Kanuri [cent2050] (Saharan; Cameroon, Niger, Nigeria, and Chad) has a DOM in which the P marker *-ga* appears under specific word orders (categorized under `<Miscellaneous_condition>`), while in the standard word order P is not marked. Brahui [brah1256] (Dravidian; Pakistan, Iran, and Afghanistan) has a DOM based on definiteness (categorized under `<Reference>`): indefinite nominal P arguments are not marked for

²³ Because the `references.csv` table does not calculate alignment according to fixed coarguments but generalizes across them (see Section 5.3), P selectors that occur for the same reference with different coarguments cooccur in a single cell. In such cases, the relevant row receives the label `<sensitive>`, indicating coargument-sensitive alignment. The code for calculating DOMs made available in the Supplementary Materials makes the assumption that all such coargument-sensitive differential P flagging necessarily implies the presence of DOM, without calculating all fixed coargument alignments. A manual check confirms that this assumption is correct, at least for the data present in our database.

Table 10: Excerpt from `doms.csv`

Glotto-code	Conditioning	Marking_1	Alignment_1	Condition_1	Marking_2	Alignment_2	Condition_2
cent2050	Misc_cond	NO_FLAGGING	no marking	default	P marker -ga	accusative	non-standard order
brah1256	Reference	ACC -e	accusative	Noun-def; Pro	NO_FLAGGING	no marking	Noun-indef

case, while definite nominal and pronominal P arguments have accusative P marking.

With this role-level aggregation, we can derive yet another language-level property, namely, whether the language has DOM at all and what the triggering condition is for the DOM. This is added to the `values.csv` table, using the `doms.csv` table to derive this information. We found that in our sample, DOM is fairly common (20% of languages), with the majority (71%) having a reference-based split.

6

CONCLUSION

When doing typological comparison on complex and multi-layered parts of grammar, such as morphosyntactic alignment, there are many possible points of comparison for the analyst to choose from. One valid method of comparison is to select a well-defined exemplar and compare languages based strictly on the exemplar case. Another possibility is to enumerate each possible pattern and ask whether each occurs in the language above a certain frequency (or whether it occurs at all). With a carefully constructed database, it is possible to encode linguistic data in a way that allows for “late aggregation” (Witzlack-Makarevich *et al.* 2022) for multiple points of comparison based on the same data structure.

We have presented such a database for alignment and shown how it can be used to answer many types of typological questions concerning core argument flagging and indexing. This includes many traditional concepts of alignment (such as alignment per referential type), broader alignment-related phenomena (such as differential object marking), and more expansive questions about argument flagging

and indexing (such as the presence of indexing at all, and which persons and roles lack indexing or are indexed by phonologically null elements). Our database is extensible and there are several additional phenomena that could be added: other roles (such as Theme and Goal); other predicate classes (beyond the major class of bivalent verbs); other types of argument selectors beyond indexing and flagging (e.g. various syntactic properties); and so on. Further aggregations of the data are also possible, besides the ones we have demonstrated. Differential agent marking, differences in alignment based on targeted features (person, person + number, or number only), and a finer distinction among zero-indexing for 3rd persons (separating by number and even gender) are some of the most obvious extensions. Beyond adding more data and more aggregations, another direction for future research could include a more streamlined user interface for data entry and quantitative comparisons with other databases of a different design philosophy. The work presented here, both in database design and ways to query data for typological properties, represents a step forward in the direction of creating generalized, multi-purpose typological databases which can be used to answer many typological questions all at once.

AUTHOR CONTRIBUTIONS

The database structure and data aggregations were conceived by A.W.M., D.I. and N.C.P. Data were collected by A.W.M. and D.I. All computer code was written by M.S. (for reference-based alignment) and D.I. (for other aggregations). The paper was written by A.W.M., D.I., and N.C.P.

ACKNOWLEDGEMENTS

We would like to thank our colleagues and research assistants who contributed to data entry: Marine Vuillermet, Anna Graff, Tai Hong, and Alexandra Nogina, as well as Balthasar Bickel for conceptual discussions. We also would like to thank our reviewers for their helpful feedback on our manuscript. D.I., N.C.P., and M.S. were supported by the Swiss National Science Foundation (SNSF) Sinergia Project “Out of Asia” CRSII5_183578.

APPENDICES

A GENERIC CLDF DATASET DESCRIPTION

The generic CLDF dataset includes a `metadata.json` file, a `sources.bib` file and five tables: `languages.csv`, `contexts.csv`, `selectors.csv`, `references.csv` and `doms.csv`. Of these, the first three tables are basic and correspond to raw data collected from grammars, while the other two are populated algorithmically through scripts. The `metadata.json` file describes the whole dataset and how the different tables are interrelated. The `sources.bib` file contains the bibliographic references. The tables are described in detail below.

A.1 *languages.csv*

Each doculect is identified through its Glottocode and its Glottolog name. This table also contains information about family membership (`Family_Name` column), macroarea, and geographic coordinates (Hammarström *et al.* 2023).

Additionally, there is a `Comment` column for any further unstructured information.

A.2 *contexts.csv*

Each context has a unique ID, and is linked to the doculect it belongs to through the `Glottocode` column and to a selector (the morpheme or paradigm of morphemes used in this context) through the `Selector_ID` column. Bibliographic references are given in the `Source` column and the responsible person in the `Coder` column. Finally, any additional remarks are kept in the `Comment` column.

The `Role` and `Reference` columns refer to the argument, while the `Co-argument_role` and `Co-argument_reference` columns to the co-argument. Note that as explained in Section 4.2.1, all contexts involving two arguments are written as two separate contexts where each argument is considered as the primary argument and the other

as the co-argument. The Role column can only take one of three values: <S> (for Sole argument of monovalent verb), <A> (for Agent-like argument of bivalent verb), or <P> (for Patient-like argument of bivalent verb). The Co-argument_role column can take only one of the following three values: <P> (when the argument is A), <A> (when the argument is P), <NA> (for Not Applicable when the argument is S). In the present form of the ATLAS Alignment Module, the Co-argument_role column is redundant since it can be predicted by the Role column. However, in an extended form of the database, where e.g. arguments of trivalent verbs are included, more combinations of argument and co-argument roles would be possible, since A could be combined with Theme or Goal.

The Reference column can take a variety of values depending on the doculect in question. For indexing, it can take any relevant person-number combination, such as <1sg>, <1pl.incl>, <3pl>, as well as any relevant gender distinction, e.g. <3sg.M>, <3sg.F>. For pronouns, the possible values are the same as for indexing for most languages, but they are always followed by the string “Pro” (e.g. <2sgPro>, <1duPro>, <3pl.F.Pro>). For nouns, the relevant categories are noun classes or other kinds of noun groups that behave uniformly in terms of alignment, always including the string “Noun” (e.g. <Noun-M>, <Noun-sg>, <Noun-pl-indef>, <Noun-high>). The Co-argument_reference column can take the same values as the Reference column, as appropriate for the co-argument restrictions of each context. The Selector_ID column always refers to the marking of the argument (rather than the co-argument) in each context. This is true even for portmanteau morphemes that mark both the A and P roles, since such morphemes appear in two different context rows, one for marking the A argument and one for marking the P argument. The Portmanteau column has also been filled out only for indexing, and indicates whether the selector involved in the context functions as a portmanteau which indexes both A and P roles; it takes three possible values: <NA>, <simple>, and <portmanteau>.

The Slot column is optional and contains information about the relative orders in which the argument markers appear. This column does not capture slot in the strict sense of a fully articulated morphological template, as determining this for every language in our large typological study proved impractical (for example, there may be

Table 11: Possible values for Slot in contexts.csv

Value	Interpretation
1, 2, etc.	suffix at the 1st, 2nd etc. slot
−1, −2, etc.	prefix at the 1st, 2nd etc. slot
0	infix or stem change (tone, ablaut, etc.)
1/ − 1	mixed paradigm that contains both prefixes and suffixes and their corresponding slots
1&2	the suffix slot could be 1 or 2 depending on the analysis
clitic 1	enclitic
clitic -1	proclitic
multiple	for Ø morphemes on verbs with multiple slots for indexing; the number of posited Ø morphemes in such cases can vary depending on the analysis
AUX −1, AUX 1, etc.	affixes at the corresponding slot on an auxiliary verb
NA	not applicable; for languages with no argument indexing

many optional slots for grammatical voice markers and TAM information that are not fully listed in the description). Instead, the value in our Slot column is only guaranteed to be correct in a relative sense: a <2> indicates a suffix further to the right of the stem than a <1>, for example. For languages with complex templatic structures, we used the slot values given in the grammar. Otherwise, we generated our own slot information based on what was present in the description of the indexing paradigm. The possible values for slot and their interpretation can be seen in Table 11.

The Exemplar column is a convenience column for the extraction of alignment patterns per referential type and contains information about our exemplar monovalent and bivalent context as explained in Section 4.1. It can take the values: <exemplar>, <non-exemplar>, and <any>. The Exemplar column value <exemplar> corresponds to cases where the context or context bundle in question fits the properties of our exemplar exactly. This value is not attested in our data, due to our exemplar being highly specified. The value <non-exemplar> indicates that the context or context bundle in question does not fit the properties of our exemplar in some regard (e.g. the A may be non-human; or it may not be in control of the action; the

action may have not happened yet; etc.). Finally, the value `<any>` indicates that this bundle of contexts can contain both exemplar and non-exemplar situations.

Non-exemplar contexts are entered as separate rows in the `contexts.csv` only if they are marked in a way that produces a different alignment pattern. Otherwise, they are bundled appropriately in corresponding `<any>` contexts.

Common conditions that cause splits in alignment and yield non-exemplar alignment patterns, such as TAM, predicate class and co-argument reference, are marked in the homonymous columns, while all other conditions are listed in the `Miscellaneous_condition` column. The TAM column can take any value following the language description, such as `<progressive>`, `<perfective>`, `<future>`, etc. The `Predicate_class` column has at least one `<default>` monovalent predicate class and one `<default>` bivalent predicate class. Beyond the default classes, languages may have any number of other predicate classes, such as `<stative verbs>`. For the present study, we have coded additional bivalent predicate classes only if they contain meanings that at least some of the time meet our exemplar conditions, as well as additional monovalent predicate classes where the S argument lacks control. This restriction is motivated by reasons of practicality (it is often difficult to find details about all predicate classes in a language and/or it takes longer to code) and because our broader study was specifically interested in the “split S” phenomenon.

Finally, several of these columns have a special value type, `<any>`, which is used as a “wildcard”: an `<any>` in the TAM, `Miscellaneous_condition`, and `Exemplar` columns signifies that the context bundle contains contexts that have all possible values of the relevant variable for this doculect. This is a way to avoid duplicate encoding of contexts which are not sensitive to conditions that may be operative in other parts of the language. For example, the language Lavukaleve [lavu1241] (isolate; Solomon Islands) sometimes drops S and A indexing on the verb in unknown discourse contexts, but always indexes P on the verb. In this case, there are contexts for S and A indexing, conditioned on `Miscellaneous_condition <default>`, and contexts for a lack of S and A indexing, conditioned on another `Miscellaneous_condition` (descriptively, `<unknown conditions, may be discourse-based>`). The indexing for P, however, occurs in

both conditions. So the P context has the `Miscellaneous_condition` `<any>`, which means that it occurs for all possible values of `Miscellaneous_condition`. The wildcard `<any>` can also be used for `Reference` and `Co-argument_Reference`, where it refers to any possible referential value. In the case where a bivalent context is not influenced by its co-argument, the value for that variable is `<any>`. In cases where all indexation or all flagging uses the same paradigm, regardless of referential properties, the `Reference` is set to `<any>`.

A.3

selectors.csv

Each selector has a unique ID and is linked to the corresponding doculect through the `Glottocode` column. Analogous to the `contexts.csv`, there are independent columns for primary reference, Source, and the coder who entered the data, `Coder`. Selectors have a name (either a high-level description or their phonological form, e.g. ‘ACC case’ or ‘-ú 3plS/A’), which is entered in the `Selector_label` column. The `Selector_type` column can take three values in our database: `<flagging>` for case marking or adpositions, `<indexing marker>` for selectors involved in verbal indexing, and `<indexing trigger>` for roles that lack verbal indexing. The `Marker_type` column is a boolean type column involving two values: `<overt>` and `<zero>`, for overt and null markers respectively. The `Features` column encodes which features a selector indexes; this column has only been filled out for indexing. It can take one of six values for our data: `<NA>`, `<person>`, `<number>`, `<person + number>`, `<gender>`, and `<other>`. The value `<other>` covers a variety of features that are more rarely attested, such as proximate/obviative, specificity, honorificity, etc.

The table also includes four columns whose values are not entered by hand, but are derived algorithmically, as described in Section 5.2: `S_references`, `A_references`, `P_references`, and `Alignment`.

A.4

references.csv

The `references.csv` is entirely derived by the `reference_alignment.py` script, the logic of which is detailed further below in Appendix C. The table lists references for every doculect

and every relevant condition and gives their alignments. Each reference has a unique ID and is linked to the corresponding doculect through the Glottocode column, and the language name is given in human-readable format in the Language column. The domain to which the reference applies is given in Domain, and has three possible values: <Noun>, <Pro> (i.e. pronoun), and <Verb>. The referential type itself is given in Referential_type and takes an open-ended set of values, which correspond to the referential types present in that language. In the Exemplar column it is indicated if the referential type and associated conditions are among the exemplar ones (value <exemplar>) or if it includes non-exemplar conditions and referential types as well (value <all>). Note that for a language that has no non-exemplar contexts (that are behaving differently from exemplar contexts as far as alignment is concerned) these sets of rows will be identical. By construction, every referent for every language will have at least one row with Exemplar marked as <all>. The value of other conditions relevant to the alignment statement is given in the columns Monovalent_predicate_class, Bivalent_predicate_class, TAM, and Condition. For languages that have multiple monovalent predicate class and/or multiple bivalent predicate classes, each monovalent predicate class is combined with each bivalent predicate class to produce alignment statements. The S, A, and P columns give the selector(s) which encode that role for each reference, and the Alignment and Alignment_not_local columns abstract over S, A, and P, generating an alignment per referent (per condition). As explained in Section 5.3, the Alignment column takes into account all scenarios in the alignment calculation, while for the Alignment_not_local columns, local scenarios are excluded. Finally, the Source column amalgamates the sources from the contexts.csv and selectors.csv tables that were used to generate this alignment, and the Coder column likewise amalgamates the coders.

doms.csv

A.5

The `doms.csv` table is entirely derived by the `dom_aggregation` function in the `alignment_aggregation.py` script, as described in Section 5.4. The table lists all DOMs (Differential Object Marking) present

in the sample, each of which has a unique ID and is linked to the corresponding doculect through the Glottocode column. The condition(s) that generate the DOM are given in the Conditioning column, which can take the values <Reference>, <Miscellaneous_condition>, <TAM>, and <Co-argument_reference>, or in the case of complex conditions, two or more of these joined by a <+>. The `doms.csv` table also includes an open-ended series of columns, `Marking_X`, `Alignment_X`, and `Condition_X`, for $X = 1, 2, \dots$, for as many conditions as there are encountered in the data for the same doculect. Each `Marking_X` column gives one of the possible markings of P, each `Alignments_X` column gives the set of alignments associated with the marking, and each `Condition_X` column gives the condition in which that marking appears. DOMs definitionally have at least two different markings under two different conditions, but in our data we have one language with three different markings following three different conditions. Finally, there is a `Source` column which amalgamates the sources in `contexts.csv` and `selectors.csv` from which this DOM was derived, and a `Coder` column which concatenates the coders.

B STRUCTURE CLDF DATASET

The structure CLDF dataset has a `metadata.json` and three tables: `languages.csv` (which is an identical copy of the one in the generic CLDF dataset), `parameters.csv`, and `values.csv`.

The `parameters.csv` table contains all language-level aggregations (including ones which are derived from selector, reference, and role-level aggregations), in the form of a unique `Parameter_ID` and a `Question`, which describes the typological property that is derived in the form of a question.

The value of a particular doculect for a particular parameter corresponds to a row in `values.csv`, and is associated with the `languages.csv` and the `parameters.csv` tables via the `Glottocode` and `Parameter_ID` columns respectively. The value itself is stored in the `Value` column. Finally, values have a `Coder`, which is the concatenation of all coders responsible for the raw data which generated this value, and a `Source`, which is the amalgamation of all sources in the

raw data which generated this value. As mentioned in Appendix 5.1, an alternative view of this information – as a matrix with one row per doculect and a column for each parameter – can be generated from our scripts and output by default to `human-readable.csv`. This file is not part of the structure CLDF dataset.

REFERENCE-LEVEL AGGREGATION CODE FLOW EXAMPLE

C

In this section, we present two examples of the code flow of the `reference_alignment.py` script, which calculates reference-based alignment, first for Kamu [kamu1258] (Kamu; Australia) and then for Marind [nucl1622] (Anim; Indonesia and Papua New Guinea). Note that the tables we present are slightly simplified with invariant or non-relevant columns removed for the sake of readability.

Kamu exemplifies a moderately complex system of both flagging and indexing, each of which can change under different conditions. Kamu has 26 rows in the `contexts.csv` table (four for argument flagging and 22 for verbal indexing, see Table 12), and five selectors in the `selectors.csv` table (two for flagging and three for indexing, see Table 13).

First, we will exemplify the calculation of alignment for referential types that receive flagging. By filtering the `contexts.csv` table for selectors which are used in flagging (whose corresponding `Selector_type` in the `selectors.csv` is `<flagging>`), we see that there is only one referential type, the special type `<any>`, indicating that all pronouns and nouns behave identically with respect to alignment, and two miscellaneous conditions (`<default>` and `<non-default>`). For each referential type (in this case, only `<any>`), we filter the table for every possible miscellaneous condition (here, `<default>` and `<non-default>`), always matching `<any>` with all other values, as explained in A.2. As an example, filtering for the referential type `<any>` and the miscellaneous condition `<default>` yields Table 14.

This resulting table is used to fill in the corresponding row for referential type `<any>` and miscellaneous condition `<default>` in the

Table 12: Kamu contexts

ID	Selector_ID	Slot	Role	Reference	Co-argument_ reference	Exemplar	Miscellaneous_ condition
kamu1258-1	kamu1258-erg-marking-flagging		A	any	any	any	default
kamu1258-2	kamu1258-no-flagging		A	any	any	any	non-default
kamu1258-3	kamu1258-no-flagging		S	any	NA	any	any
kamu1258-4	kamu1258-no-flagging		P	any	any	any	any
kamu1258-5	kamu1258-null-marker	1	P	3sg.nonhum	any	non-exemplar	any
kamu1258-6	kamu1258-null-marker	1	P	3sg.hum	any	any	unknown condition
kamu1258-7	kamu1258-p-enclitics-indexing-marker	1	P	1sg	any	any	any
kamu1258-8	kamu1258-p-enclitics-indexing-marker	1	P	1pl	any	any	any
kamu1258-9	kamu1258-p-enclitics-indexing-marker	1	P	2sg	any	any	any
kamu1258-10	kamu1258-p-enclitics-indexing-marker	1	P	2pl	any	any	any
kamu1258-11	kamu1258-p-enclitics-indexing-marker	1	P	3pl	any	any	any
kamu1258-12	kamu1258-p-enclitics-indexing-marker	1	P	3sg.hum	any	any	default
kamu1258-13	kamu1258-s-a-prefixes-indexing-marker	AUX -1	A	1sg	any	any	any
kamu1258-14	kamu1258-s-a-prefixes-indexing-marker	AUX -1	A	2sg	any	any	any
kamu1258-15	kamu1258-s-a-prefixes-indexing-marker	AUX -1	A	2pl	any	any	any
kamu1258-16	kamu1258-s-a-prefixes-indexing-marker	AUX -1	A	1pl	any	any	any
kamu1258-17	kamu1258-s-a-prefixes-indexing-marker	AUX -1	A	3pl	any	any	any
kamu1258-18	kamu1258-s-a-prefixes-indexing-marker	AUX -1	A	3sg.nonhum	any	non-exemplar	any
kamu1258-19	kamu1258-s-a-prefixes-indexing-marker	AUX -1	A	3sg.hum	any	any	any
kamu1258-20	kamu1258-s-a-prefixes-indexing-marker	AUX -1	S	1sg	NA	any	any
kamu1258-21	kamu1258-s-a-prefixes-indexing-marker	AUX -1	S	1pl	NA	any	any
kamu1258-22	kamu1258-s-a-prefixes-indexing-marker	AUX -1	S	2sg	NA	any	any
kamu1258-23	kamu1258-s-a-prefixes-indexing-marker	AUX -1	S	2pl	NA	any	any
kamu1258-24	kamu1258-s-a-prefixes-indexing-marker	AUX -1	S	3pl	NA	any	any
kamu1258-25	kamu1258-s-a-prefixes-indexing-marker	AUX -1	S	3sg.nonhum	NA	non-exemplar	any
kamu1258-26	kamu1258-s-a-prefixes-indexing-marker	AUX -1	S	3sg.hum	NA	any	any

Table 13: Kamu selectors

ID	Selector_type	Selector_label	Marker_type	Features
kamu1258-erg-marking-flagging	flagging	ERG marking	overt	
kamu1258-no-flagging	flagging	NO_FLAGGING	zero	
kamu1258-null-marker	indexing marker	NULL_MARKER	zero	NA
kamu1258-p-enclitics-indexing-marker	indexing marker	P enclitics	overt	person + number
kamu1258-s-a-prefixes-indexing-marker	indexing marker	S/A prefixes	overt	person + number

Table 14: Kamu contexts: filtering for selector type < flagging > and < default > condition

ID	Selector_ID	Role	Reference	Co-argument_ reference	Exemplar	Miscellaneous_ condition
kamu1258-1	kamu1258-erg-marking-flagging	A	any	any	any	default
kamu1258-3	kamu1258-no-flagging	S	any	NA	any	any
kamu1258-4	kamu1258-no-flagging	P	any	any	any	any

`references.csv` table as follows: The column S contains the selector (and if it is overt or not) for referential type `<any>` when in the S role, and the same for columns A and P. The result of this process is shown in the first row of Table 15. In the same way, now filtering for `<any>` and `<non-default>` condition, we can fill in the second row of Table 15. Note that there are two sets of rows in the `references.csv` table: one set of rows is marked `<exemplar>` in the Exemplar column and includes only exemplar contexts, and the other is marked `<all>` and includes all contexts (exemplar and non-exemplar). In a subsequent step, the S, A, P columns of `references.csv` are used to calculate the alignment for each referential type. Here, all nouns and pronouns (referential type `<any>`) have an ergative alignment in the `<default>` condition, since only the A argument is marked with an overt marker. In the `<non-default>` condition all nouns and pronouns receive no marking since all selectors are of Marker_type `<zero>`.

Indexing in Kamu changes depending on referential properties that are included in our exemplar, with some referential types conditionally marked by a null morpheme. We again filter for each combination of referential type, exemplar case, and any relevant miscellaneous condition (i.e. a condition within a certain exemplar case). In this case, Exemplar `<non-exemplar>` always has Miscellaneous_condition `<any>`, and Exemplar `<any>` has `<unknown condition>`, `<default>` or `<any>`, so the following combinations are filtered for in different iterations of the script: each person-number combination for Exemplar `<any>` and Miscellaneous_condition `<unknown condition>` or `<any>`, and each person-number combination for Exemplar `<any>` and Miscellaneous_condition `<default>` or `<any>`. During each iteration, a corresponding line is filled in `references.csv`, following the same process as for flagging. In the case of Kamu indexing, all referential types have an accusative alignment, even though the P marking changes under certain conditions. The full reference-based alignment table for both flagging and indexing in Kamu, after all processing is done, is presented in Table 15.

Our other example, Marind, has no flagging at all, but a different kind of complexity in its indexing system, including both a split in S marking according to predicate class and co-argument sensitivity for 3pl. Contexts for Marind are given in Table 16 (22 rows total) and selectors in Table 17 (nine rows total).

Table 15: Kamu reference-based alignments

Domain	Referen- tial_type	Exemplar	Condition	S	A	P	Alignment	Alignment_ not_local
Noun	any	exemplar	default	NO_FLAGGING_zero	ERG marking_overt	NO_FLAGGING_zero	ergative	ergative
Noun	any	exemplar	non-default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
Noun	any	all	default	NO_FLAGGING_zero	ERG marking_overt	NO_FLAGGING_zero	ergative	ergative
Noun	any	all	non-default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
Pro	any	exemplar	non-default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
Pro	any	exemplar	default	NO_FLAGGING_zero	ERG marking_overt	NO_FLAGGING_zero	ergative	ergative
Pro	any	all	non-default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking	no marking
Pro	any	all	default	NO_FLAGGING_zero	ERG marking_overt	NO_FLAGGING_zero	ergative	ergative
Verb	2pl	exemplar	unknown condition	S/A prefixes_overt	ERG marking_overt	P enditics_overt	accusative	accusative
Verb	2sg	exemplar	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	1pl	exemplar	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	3sg.hum	exemplar	unknown condition	S/A prefixes_overt	S/A prefixes_overt	NULL_MARKER_zero	accusative	accusative
Verb	3pl	exemplar	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	1sg	exemplar	default	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	2pl	exemplar	default	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	2sg	exemplar	default	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	1pl	exemplar	default	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	3sg.hum	exemplar	default	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	3pl	exemplar	default	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	1sg	all	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	2pl	all	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	2sg	all	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	1pl	all	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	3sg.nonhum	all	unknown condition	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	3sg.hum	all	unknown condition	S/A prefixes_overt	S/A prefixes_overt	NULL_MARKER_zero	accusative	accusative
Verb	3pl	all	unknown condition	S/A prefixes_overt	S/A prefixes_overt	NULL_MARKER_zero	accusative	accusative
Verb	1sg	all	default	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	2pl	all	default	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	2sg	all	default	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	1pl	all	default	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	3sg.nonhum	all	default	S/A prefixes_overt	S/A prefixes_overt	NULL_MARKER_zero	accusative	accusative
Verb	3sg.hum	all	default	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative
Verb	3pl	all	default	S/A prefixes_overt	S/A prefixes_overt	P enditics_overt	accusative	accusative

Table 16: Marind contexts

ID	Selector_ID	Slot	Role	Reference	Co-argument_ reference	Exemplar	Predicate_ class
nucl1622-1	nucl1622-no-flagging		A	any	any	any	default
nucl1622-2	nucl1622-no-flagging		P	any	any	any	default
nucl1622-3	nucl1622-no-flagging		S	any	NA	any	default
nucl1622-4	nucl1622-p-affix-indexing-marker	-1/1	P	any	any	any	default
nucl1622-5	nucl1622-s-a-prefix-1pl-nak-e-indexing-marker	-1/1	S	any	NA	non-exemplar	Sp class
nucl1622-6	nucl1622-s-a-prefix-1pl-nak-e-indexing-marker	-1	S	1pl	NA	any	default
nucl1622-7	nucl1622-s-a-prefix-1sg-no-nak-indexing-marker	-1	A	1pl	any	any	default
nucl1622-8	nucl1622-s-a-prefix-1sg-no-nak-indexing-marker	-1	S	1sg	NA	any	default
nucl1622-9	nucl1622-s-a-prefix-2pl-e-indexing-marker	-1	A	1sg	any	any	default
nucl1622-10	nucl1622-s-a-prefix-2pl-e-indexing-marker	-1	S	2pl	NA	any	default
nucl1622-11	nucl1622-s-a-prefix-2sg-o-indexing-marker	-1	A	2pl	any	any	default
nucl1622-12	nucl1622-s-a-prefix-2sg-o-indexing-marker	-1	S	2sg	NA	any	default
nucl1622-13	nucl1622-p-affix-indexing-marker	-1	A	2sg	any	any	default
nucl1622-14	nucl1622-s-a-prefix-3pl-e-3pl-1-indexing-marker	-1	A	3pl	1sg	any	default
nucl1622-15	nucl1622-s-a-prefix-3pl-e-3pl-1-indexing-marker	-1	A	3pl	1pl	any	default
nucl1622-16	nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	-1	S	3pl	NA	any	default
nucl1622-17	nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	-1	A	3pl	2sg	any	default
nucl1622-18	nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	-1	A	3pl	2pl	any	default
nucl1622-19	nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	-1	A	3pl	3sg	any	default
nucl1622-20	nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	-1	A	3pl	3pl	any	default
nucl1622-21	nucl1622-s-a-prefix-3sg-a-0-indexing-marker	-1	S	3sg	NA	any	default
nucl1622-22	nucl1622-s-a-prefix-3sg-a-0-indexing-marker	-1	A	3sg	any	any	default

Table 17: Marind selectors

ID	Selector_type	Selector_label	Marker_type	Features
nucl1622-no-flagging	flagging	NO_FLAGGING	zero	
nucl1622-p-affix-indexing-marker	indexing marker	P affix	overt	person + number
nucl1622-s-a-prefix-1pl-nak-e-indexing-marker	indexing marker	S/A prefix 1pl nak-...(e-)	overt	person + number
nucl1622-s-a-prefix-1sg-no-nak-indexing-marker	indexing marker	S/A prefix 1sg no-/nak-	overt	person + number
nucl1622-s-a-prefix-2pl-e-indexing-marker	indexing marker	S/A prefix 2pl e-	overt	person + number
nucl1622-s-a-prefix-2sg-o-indexing-marker	indexing marker	S/A prefix 2sg o-	overt	person + number
nucl1622-s-a-prefix-3pl-e-3pl-1-indexing-marker	indexing marker	S/A prefix 3pl e- (3pl > 1)	overt	person + number
nucl1622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	indexing marker	S/A prefix 3pl n- (not in 3pl > 1)	overt	person + number
nucl1622-s-a-prefix-3sg-a-0-indexing-marker	indexing marker	S/A prefix 3sg a-/0-	overt	person + number

Table 18: Marind contexts filtered for <default> monovalent predicate class and <1pl>

ID	Selector_ID	Slot	Role	Reference	Co-argument_reference	Predicate class
nucl1622-4	nucl1622-p-affix-indexing-marker	-1/1	P	any	any	default
nucl1622-6	nucl1622-s-a-prefix-1pl-nak-e-indexing-marker	-1	S	1pl	NA	default
nucl1622-7	nucl1622-s-a-prefix-1pl-nak-e-indexing-marker	-1	A	1pl	any	default

Marind has two monovalent predicate classes (<default> and <Sp class>), so there are two different alignment calculations: one comparing monovalent default S with bivalent default A and P and one comparing monovalent Sp class S with bivalent default A and P. An alignment is calculated for each referential type and for each condition (in this case, default and Sp class). Filtering for the <1pl> referential type and the <default> monovalent predicate class results in Table 18. Note that referential type <any> matches all specific referential types, including <1pl>. The alignment for <1pl> and the <default> condition is accusative as can be seen in the fifth row of Table 20 for the set of exemplar alignments and in the 17th row for the set of all alignments. Note that in the set of all alignments, an additional alignment statement for <1pl> is attested (in the eleventh row of the table), this time with a different monovalent predicate class (Sp) and its alignment value is <ergative>. Predicates of the Sp class indicate actions where the S has no control, and therefore they are not included in the set of exemplar alignments, since our chosen exemplar requires that the S has control over the event (see Section 3).

When co-argument sensitivity is involved, a referential type will participate in multiple contexts with the same role but with different co-arguments, as is the case for 3pl in Marind. This can be seen in Table 19, which filters Marind contexts for referential type <3pl> and the <default> monovalent predicate class. An alignment for this referential type cannot be calculated because there is no single marker for the A role (although one could calculate an alignment if the co-arguments were fixed, i.e. the marking of 3pl when its co-argument, if any, is 1sg – e.g. an alignment of 3pl S vs. A (with 1sg P) vs. P (with 1sg A) – but this is not something we have done here). Instead, in *references.csv* all the different ways that 3pl A is marked depending on the co-argument are concatenated within the same cell of the A column (see the first, seventh and thirteenth rows in Table 20). When we calculate reference-based alignments, cases such as 3pl in Marind get the pseudo-alignment <sensitive>, indicating that there is no single alignment statement that can be made without the co-argument role being fixed.

Once all of these calculations are done for every reference and every condition, the output for reference-based alignment of indexing in Marind is as in Table 20.

Table 19: Marind contexts filtered for <default> monovalent predicate class and <3pl>

ID	Selector_ID	Slot	Role	Reference	Co-argument reference	Predicate class
nucl1622-4	nucl1 622-p-affix-indexing-marker	-1/1	P	any	any	default
nucl1622-14	nucl1 622-s-a-prefix-3pl-e-3pl-1-indexing-marker	-1	A	3pl	1sg	default
nucl1622-15	nucl1 622-s-a-prefix-3pl-e-3pl-1-indexing-marker	-1	A	3pl	1pl	default
nucl1622-16	nucl1 622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	-1	S	3pl	NA	default
nucl1622-17	nucl1 622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	-1	A	3pl	2sg	default
nucl1622-18	nucl1 622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	-1	A	3pl	2pl	default
nucl1622-19	nucl1 622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	-1	A	3pl	3sg	default
nucl1622-20	nucl1 622-s-a-prefix-3pl-n-not-in-3pl-1-indexing-marker	-1	A	3pl	3pl	default

Table 20: Marind reference-based alignments

Referential_ type	Exemplar	Monovalent_ predicate_class	S	A	P	Alignment
3pl	exemplar	default	S/A prefix 3pl n- (not in 3pl>1)_overt	S/A prefix 3pl e- (3pl>1)_overt_coarg:1sg ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:2sg ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:2pl ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:3sg ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:3pl ; INFERRED_NULL_zero_coarg:else	P affix_overt	sensitive
2pl	exemplar	default	S/A prefix 2pl e-_overt	S/A prefix 2pl e-_overt	P affix_overt	accusative
1sg	exemplar	default	S/A prefix 1sg no-/nak-_overt	S/A prefix 1sg no-/nak-_overt	P affix_overt	accusative
2sg	exemplar	default	S/A prefix 2sg o-_overt	S/A prefix 2sg o-_overt	P affix_overt	accusative
1pl	exemplar	default	S/A prefix 1pl nak-...(e-)_overt	S/A prefix 1pl nak-...(e-)_overt	P affix_overt	accusative
3sg	exemplar	default	S/A prefix 3sg a-/0-_overt	S/A prefix 3sg a-/0-_overt	P affix_overt	accusative
3pl	all	Sp class	P affix_overt	S/A prefix 3pl e- (3pl>1)_overt_coarg:1sg ; S/A prefix 3pl e- (3pl>1)_overt_coarg:1pl ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:2sg ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:2pl ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:3sg ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:3pl ; INFERRED_NULL_zero_coarg:else	P affix_overt	sensitive
2pl	all	Sp class	P affix_overt	S/A prefix 2pl e-_overt	P affix_overt	ergative
1sg	all	Sp class	P affix_overt	S/A prefix 1sg no-/nak-_overt	P affix_overt	ergative
2sg	all	Sp class	P affix_overt	S/A prefix 2sg o-_overt	P affix_overt	ergative
1pl	all	Sp class	P affix_overt	S/A prefix 1pl nak-...(e-)_overt	P affix_overt	ergative
3sg	all	Sp class	P affix_overt	S/A prefix 3sg a-/0-_overt	P affix_overt	ergative
3pl	all	default	S/A prefix 3pl n- (not in 3pl>1)_overt	S/A prefix 3pl e- (3pl>1)_overt_coarg:1sg ; S/A prefix 3pl e- (3pl>1)_overt_coarg:1pl ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:2sg ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:2pl ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:3sg ; S/A prefix 3pl n- (not in 3pl>1)_overt_coarg:3pl ; INFERRED_NULL_zero_coarg:else	P affix_overt	sensitive
2pl	all	default	S/A prefix 2pl e-_overt	S/A prefix 2pl e-_overt	P affix_overt	accusative
1sg	all	default	S/A prefix 1sg no-/nak-_overt	S/A prefix 1sg no-/nak-_overt	P affix_overt	accusative
2sg	all	default	S/A prefix 2sg o-_overt	S/A prefix 2sg o-_overt	P affix_overt	accusative
1pl	all	default	S/A prefix 1pl nak-...(e-)_overt	S/A prefix 1pl nak-...(e-)_overt	P affix_overt	accusative
3sg	all	default	S/A prefix 3sg a-/0-_overt	S/A prefix 3sg a-/0-_overt	P affix_overt	accusative
any	exemplar	default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking
any	all	default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking
any	exemplar	default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking
any	all	default	NO_FLAGGING_zero	NO_FLAGGING_zero	NO_FLAGGING_zero	no marking

REFERENCES

- Stephen R. ANDERSON (1976), On the notion of subject in ergative languages, in Charles N. LI, editor, *Subject and topic*, pp. 1–23, Academic Press, New York.
- Balthasar BICKEL (2011), Grammatical relations typology, in Jae Jung SONG, editor, *The Oxford handbook of linguistic typology*, pp. 399–444, Oxford University Press, Oxford.
- Balthasar BICKEL, Giorgio IEMMOLO, Taras ZAKHARKO, and Alena WITZLACK-MAKAREVICH (2013), Patterns of alignment in verb agreement, in Dik BAKKER and Martin HASPELMATH, editors, *Languages across boundaries: Studies in memory of Anna Siewierska*, pp. 15–36, De Gruyter Mouton, Berlin.
- Balthasar BICKEL and Johanna NICHOLS (2002), Autotypologizing databases and their use in fieldwork, in Peter AUSTIN, Helen DRY, and Peter WITTENBURG, editors, *Proceedings of the International LREC Workshop on Resources and Tools in Field Linguistics, Las Palmas, 26–27 May 2002*, MPI for Psycholinguistics, Nijmegen.
- Balthasar BICKEL, Johanna NICHOLS, Taras ZAKHARKO, Alena WITZLACK-MAKAREVICH, Kristine HILDEBRANDT, Michael RIESSLER, Lennart BIERKANDT, Fernando ZÚÑIGA, and John B. LOWE (2022), The AUTOTYP database (v1.1.0), doi:10.5281/zenodo.6793367.
- Balthasar BICKEL, Alena WITZLACK-MAKAREVICH, Kamal K. CHOUDHARY, Matthias SCHLESEWSKY, and Ina BORNKESSEL-SCHLESEWSKY (2015a), The neurophysiology of language processing shapes the evolution of grammar: Evidence from case marking, *PloS One*, 10(8):e0132819.
- Balthasar BICKEL, Alena WITZLACK-MAKAREVICH, and Taras ZAKHARKO (2015b), Typological evidence against universal effects of referential scales on case alignment, in Ina BORNKESSEL-SCHLESEWSKY, Andrej L. MALCHUKOV, and Marc RICHARDS, editors, *Scales and hierarchies: A cross-disciplinary perspective*, pp. 7–43, de Gruyter Mouton, Berlin.
- Balthasar BICKEL, Alena WITZLACK-MAKAREVICH, Taras ZAKHARKO, and Giorgio IEMMOLO (2015c), Exploring diachronic universals of agreement: Alignment patterns and zero marking across person categories, in Jürg FLEISCHER, Elisabeth RIEKEN, and Paul WIDMER, editors, *Agreement from a diachronic perspective*, pp. 29–52, de Gruyter Mouton, Berlin.
- Balthasar BICKEL, Taras ZAKHARKO, Lennart BIERKANDT, and Alena WITZLACK-MAKAREVICH (2014), Semantic role clustering: An empirical assessment of semantic role types in non-default case assignment, *Studies in Language*, 38(3):485–511.

- Joshua BIRCHALL (2014a), Argument marking (argex), in Harald HAMMARSTRÖM, Olga KRASNOUKHOVA, Neele MÜLLER, Joshua BIRCHALL, Simon VAN DE KERKE, Loretta O'CONNOR, Swintha DANIELSEN, Rik VAN GIJN, and George SAAD, editors, *South American Indian language structures (SAILS) online*, Max Planck Institute for the Science of Human History, <http://sails.clld.org>.
- Joshua Thomas Rigo BIRCHALL (2014b), *Argument marking patterns in South American languages*, Utrecht: LOT.
- Georg BOSSONG (1985), *Empirische Universalienforschung: Differentielle Objektmarkierung in neuiranischen Sprachen [Empirical research on universals: Differential object marking in New Iranian languages]*, Narr, Tübingen.
- Georg BOSSONG (1991), Differential object marking in Romance and beyond, in Dieter WANNER and Douglas A. KIBBEE, editors, *New analyses in Romance linguistics. Selected papers from the XVIII Linguistic Symposium on Romance Languages Urbana-Champaign, April 7–9, 1988*, pp. 143–170, John Benjamins, Amsterdam.
- Bernard COMRIE (1978), Ergativity, in Winfred Philipp LEHMANN, editor, *Syntactic typology: Studies in the phenomenology of language*, pp. 329–394, University of Texas Press, Austin.
- Bernard COMRIE (1989), *Language universals and linguistic typology*, Blackwell, Oxford.
- Bernard COMRIE (2013a), Alignment of case marking of full noun phrases, in Matthew S. DRYER and Martin HASPELMATH, editors, *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig, <https://wals.info/chapter/98>.
- Bernard COMRIE (2013b), Alignment of case marking of pronouns, in Matthew S. DRYER and Martin HASPELMATH, editors, *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig, <https://wals.info/chapter/99>.
- Bernard COMRIE, Martin HASPELMATH, and Balthasar BICKEL (2008), The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses, *Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology & the Department of Linguistics of the University of Leipzig*.
- Sonia CRISTOFARO, Vittorio GANFI, and Guglielmo INGLESSE, editors (2021), *The Pavia DEMa (Diachronic Emergence of Alignment) database*, Università di Pavia, Pavia, <https://su-lab.unipv.it/tasf/index.php/dema/>.
- William CROFT (2001), *Radical Construction Grammar: Syntactic theory in typological perspective*, Oxford University Press, Oxford.
- Timothy CURNOW (1997), *A grammar of Awa Pit (Cuaiquer): An indigenous language of south-western Colombia*, Ph.D. thesis, Australian National University, <http://monolith.eva.mpg.de/~haspelmt/AwaPit.pdf>.

- Jaime DE ANGULO and Lucy S. FREELAND (1930), The Achumawi language, *International Journal of American Linguistics*, 6(2):77–120.
- Scott DELANCEY (2021), Differential innovation in 2nd person pronouns and agreement indexation in Trans-Himalayan languages, *Folia Linguistica*, 55(s42-s1):155–174, doi:10.1515/flin-2021-2017.
- Robert M. W. DIXON (1994), *Ergativity*, Cambridge University Press, Cambridge.
- David R. DOWTY (1991), Thematic proto-roles and argument selection, *Language*, 67(3):547–619.
- Matthew S. DRYER (1996), *Grammatical relations in Kutenai*, Voices of Rupert's Land, Winnipeg.
- Matthew S. DRYER (1997), Are grammatical relations universal?, in Joan BYBEE, John HAIMAN, and Sandra A. THOMPSON, editors, *Essays on language function and language type: Dedicated to T. Givón*, pp. 115–143, Benjamins, Amsterdam.
- Matthew S. DRYER and Martin HASPELMATH, editors (2013), *The World Atlas of Language Structures Online (v2020.3)*, Max Planck Institute for Evolutionary Anthropology, doi:10.5281/zenodo.7385533.
- Barbara Wedemeyer EDMONSON (1988), *A descriptive grammar of Huastec (Potosino dialect)*, Ph.D. thesis, Tulane University, Ann Arbor.
- Robert FORKEL, Johann-Mattis LIST, Simon J. GREENHILL, Christoph RZYMSKI, Sebastian BANK, Michael CYSOUW, Harald HAMMARSTRÖM, Martin HASPELMATH, Gereon A. KAIPING, and Russell D. GRAY (2018), Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics, *Scientific Data*, 5(1):180–205, doi:10.1038/sdata.2018.205.
- Jesús Mario GIRÓN HIGUITA (2008), *Una gramática del Wansöjöt (Puinave)*, Ph.D. thesis, University of Amsterdam.
- Harald HAMMARSTRÖM and Mark DONOHUE (2014), Some principles on the use of macro-areas in typological comparison, *Language Dynamics and Change*, 4(1):167–187, doi:10.1163/22105832-00401001.
- Harald HAMMARSTRÖM, Robert FORKEL, Martin HASPELMATH, and Sebastian BANK (2023), glottolog/glottolog-cldf: Glottolog database 4.8 as CLDF, doi:10.5281/zenodo.8131091.
- Iren HARTMANN, Martin HASPELMATH, and Bradley TAYLOR, editors (2013), *The Valency Patterns Leipzig online database*, Max Planck Institute for Evolutionary Anthropology, Leipzig, <https://valpal.info/>.
- Martin HASPELMATH (2011), On S, A, P, T, and R as comparative concepts for alignment typology, *Linguistic Typology*, 15(3):535–567.

Jeffrey HEATH (1998), Pragmatic skewing in 1 > 2 pronominal combinations in Native American languages, *International Journal of American Linguistics*, 64:83–104.

David INMAN, Natalia CHOUSOU-POLYDOURI, Marine VUILLERMET, Kellen Parker VAN DAM, Shelece EASTERDAY, Françoise ROSE, Alena WITZLACK-MAKAREVICH, Kevin M BÄTSCHER, Oscar COCAUD-DEGRÈVE, Anna GRAFF, Selma HARDEGGER, Tai HONG, Thomas C HUBER, Diana KRASOVSKAYA, Raphaël LUFFROY, Nora MUHEIM, André MÜLLER, Alexandra NOGINA, David Timothy PERROT, and Balthasar BICKEL (in prep), The ATLAS database: Areal typology of the languages of the Americas.

Edward L. KEENAN (1976), Remarkable subjects in Malagasy, in Charles LI, editor, *Subject and Topic*, Academic Press, New York.

Randy J. LAPOLLA (1993), Arguments against ‘subject’ and ‘object’ as viable concepts in Chinese, *Bulletin of the Institute of History and Philology, Academia Sinica*, 63:759–813.

Gilbert LAZARD (2002), Transitivity revisited as an example of a more strict approach in typological research, *Folia Linguistica*, 36(3–4):141–190, doi:10.1515/flin.2002.36.3-4.141.

Charles N. LI and Sandra A. THOMPSON (1976), Subject and topic: a new typology of language, in Charles N. LI, editor, *Subject and topic*, New York.

Andrej MALCHUKOV, Martin HASPELMATH, and Bernard COMRIE (2010a), Ditransitive constructions: A typological overview, in Andrej MALCHUKOV, Martin HASPELMATH, and Bernard COMRIE, editors, *Studies in ditransitive constructions: A comparative handbook*, pp. 1–35, De Gruyter Mouton, Berlin.

Andrej MALCHUKOV, Martin HASPELMATH, and Bernard COMRIE, editors (2010b), *Studies in ditransitive constructions: A comparative handbook*, De Gruyter Mouton, Berlin.

Graham MALLINSON and Barry BLAKE (1981), *Language typology: Cross-linguistic studies in syntax*, North-Holland, Amsterdam.

Edith MORAVCSIK (1978), On the distribution of ergative and accusative patterns, *Lingua*, 45(3–4):233–279.

Johanna NICHOLS (1992), *Linguistic diversity in space and time*, University of Chicago Press, Chicago.

Kazuko OBATA (2003), *A grammar of Bilua: A Papuan language of the Solomon Islands*, Research School of Pacific and Asian Studies, Australian National University, Canberra.

Simon OVERALL (2017), *A grammar of Aguaruna (Iiniá Chicham)*, De Gruyter Mouton, Berlin.

Beatrice PRIMUS (1999), *Cases and thematic roles*, Niemeyer, Tübingen.

Beatrice PRIMUS (2006), Mismatches in semantic role hierarchies and the dimensions of role semantics, in Ina BORNKESSEL, Matthias SCHLESEWSKY, Bernard COMRIE, and Angela D. FRIEDERICI, editors, *Semantic role universals and argument linking: Theoretical, typological and psycholinguistic perspectives*, pp. 53–88, Mouton de Gruyter, Berlin.

Sergey SAY, editor (2020–), *BivalTyp: Typological database of bivalent verbs and their encoding frames*, <https://www.bivaltyp.info>.

Anna SIEWIERSKA (1998), On nominal and verbal person marking, *Linguistic Typology*, 2:1–55.

Anna SIEWIERSKA (2003), Person agreement and the determination of alignment, *Transactions of the Philological Society*, 101:339–370.

Anna SIEWIERSKA (2013a), Alignment of verbal person marking, in Matthew S. DRYER and Martin HASPELMATH, editors, *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig, <http://wals.info/chapter/100>.

Anna SIEWIERSKA (2013b), Verbal person marking (v2020.3), in Matthew S. DRYER and Martin HASPELMATH, editors, *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig, <http://wals.info/chapter/102>.

Michael SILVERSTEIN (1976), Hierarchy of features and ergativity, in ROBERT M. W. DIXON, editor, *Grammatical categories in Australian languages*, pp. 112–171, Humanities Press, New Jersey.

Hedvig SKIRGÅRD, Hannah J. HAYNIE, Damián E. BLASI, Harald HAMMARSTRÖM, Jeremy COLLINS, Jay J. LATARCHE, Jakob LESAGE, Tobias WEBER, Alena WITZLACK-MAKAREVICH, Sam PASSMORE, Angela CHIRA, Luke MAURITS, Russell DINNAGE, Michael DUNN, Ger REESINK, Ruth SINGER, Claire BOWERN, Patience EPPS, Jane HILL, Outi VESAKOSKI, Martine ROBBEETS, Noor Karolin ABBAS, Daniel AUER, Nancy A. BAKKER, Giulia BARBOS, Robert D. BORGES, Swintha DANIELSEN, Luise DORENBUSCH, Ella DORN, John ELLIOTT, Giada FALCONE, Jana FISCHER, Yustinus GHANGGO ATE, Hannah GIBSON, Hans-Philipp GÖBEL, Jemima A. GOODALL, Victoria GRUNER, Andrew HARVEY, Rebekah HAYES, Leonard HEER, Roberto E. HERRERA MIRANDA, Nataliia HÜBLER, Biu HUNTINGTON-RAINEY, Jessica K. IVANI, Marilen JOHNS, Erika JUST, Eri KASHIMA, Carolina KIPF, Janina V. KLINGENBERG, Nikita KÖNIG, Aikaterina KOTI, Richard G. A. KOWALIK, Olga KRASNOUKHOVA, Nora L.M. LINDVALL, Mandy LORENZEN, Hannah LUTZENBERGER, Tônia R.A. MARTINS, Celia MATA GERMAN, Suzanne VAN DER MEER, Jaime MONTOYA SAMAMÉ, Michael MÜLLER, Saliha MURADOGLU, Kelsey NEELY, Johanna NICKEL, Miina NORVIK, Cheryl Akinyi OLUOCH, Jesse PEACOCK, India O.C. PEAREY, Naomi PECK, Stephanie PETIT, Sören PIEPER, Mariana POBLETE, Daniel PRESTIPINO, Linda RAABE, Amna

RAJA, Janis REIMRINGER, Sydney C. REY, Julia RIZAUEW, Eloisa RUPPERT, Kim K. SALMON, Jill SAMMET, Rhiannon SCHEMBRI, Lars SCHLABBACH, Frederick W.P. SCHMIDT, Amalia SKILTON, Wikaliler Daniel SMITH, Hilário DE SOUSA, Kristin SVERREDAL, Daniel VALLE, Javier VERA, Judith VOSS, Tim WITTE, Henry WU, Stephanie YAM, Jingting YE, Maisie YONG, Tessa YUDITHA, Roberto ZARIQUEY, Robert FORKEL, Nicholas EVANS, Stephen C. LEVINSON, Martin HASPELMATH, Simon J. GREENHILL, Quentin D. ATKINSON, and Russell D. GRAY (2023), Grambank reveals global patterns in the structural diversity of the world's languages, *Science Advances*, 9, doi:10.1126/sciadv.adg6175.

Robert D. VAN VALIN, Jr. (1981), Grammatical relations in ergative languages, *Studies in Language*, 5(3):361–394.

Robert D. VAN VALIN, Jr. (1983), Pragmatics, ergativity and grammatical relations, *Journal of Pragmatics*, 7(1):63–88.

Robert D. VAN VALIN, Jr. (2005), *Exploring the syntax-semantics interface*, Cambridge University Press, Cambridge.

Alena WITZLACK-MAKAREVICH (2011), *Typological variation in grammatical relations*, Ph.D. thesis, University of Leipzig, Leipzig.

Alena WITZLACK-MAKAREVICH (2019), Argument selectors: A new perspective on grammatical relations. An introduction, in Alena WITZLACK-MAKAREVICH and Balthasar BICKEL, editors, *Argument Selectors: A new perspective on grammatical relations*, pp. 1–38, John Benjamins, Amsterdam.

Alena WITZLACK-MAKAREVICH, Johanna NICHOLS, Kristine A. HILDEBRANDT, Taras ZAKHARKO, and Balthasar BICKEL (2022), Managing AUTOTYP data: Design principles and implementation, in Andrea L. BEREZ-KROEKER, Bradley McDONNELL, Eve KOLLER, and Lauren B. COLLISTER, editors, *The Open Handbook of Linguistic Data Management*, pp. 631–642, The MIT Press.

Alena WITZLACK-MAKAREVICH and Ilja A. SERŽANT (2018), Differential argument marking: Patterns of variation, in Ilja A. SERŽANT and Alena WITZLACK-MAKAREVICH, editors, *Diachrony of differential argument marking*, pp. 1–40, Language Science Press, Berlin.

Alena WITZLACK-MAKAREVICH, Taras ZAKHARKO, Lennart BIERKANDT, Fernando ZÚÑIGA, and Balthasar BICKEL (2016), Decomposing hierarchical alignment: co-arguments as conditions on alignment and the limits of referential hierarchies as explanations in verb agreement, *Linguistics*, 54(3):531–561.

Fernando ZÚÑIGA (2006), *Deixis and alignment: inverse systems in indigenous languages of the Americas*, John Benjamins, Amsterdam.

David Inman

ORCID 0000-0003-1892-591X
david.inman@uzh.ch

Department of Comparative Language
Science & Center for the
Interdisciplinary Study of Language
Evolution
University of Zurich

Alena Witzlack-Makarevich

ORCID 0000-0003-0138-4635
awitzlack@maoil.huji.ac.il

Hebrew University of Jerusalem

Natalia Chousou-Polydouri

ORCID 0000-0002-5693-975X
nchousoupolydouri@gmail.com

Department of Comparative Language
Science & Center for the
Interdisciplinary Study of Language
Evolution
University of Zurich

Melvin Steiger

steigermelvin@gmail.com
ORCID 0000-0001-7300-0704

Department of Informatics
University of Zurich

David Inman, Alena Witzlack-Makarevich, Natalia Chousou-Polydouri,
and Melvin Steiger These authors have contributed equally to this work.
(2024), *Alignment everywhere all at once: Applying the late aggregation principle
to a typological database of argument marking*, *Journal of Language Modelling*,
12(2):287–347

DOI <https://dx.doi.org/10.15398/jlm.v12i2.360>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

CC BY <http://creativecommons.org/licenses/by/4.0/>

Zero marking in inflection: A token-based approach

Laura Becker
University of Freiburg

ABSTRACT

This study examines zero marking, i.e. the absence of an overt exponent, in adjectival, nominal, and verbal inflectional morphology across languages. The first part of the study provides an overview of the distribution of zero markers in inflection paradigms using the UniMorph dataset. The results show that there is a general preference against zero marking. The distribution of zero markers varies to a great extent across languages and lemmas, the only robust trend being that they are avoided in cells that express a high number of grammatical values. The second part of this study examines the association between marker frequencies and phonological length, using the Universal Dependencies treebanks. While token frequency is a good predictor for the length of overt markers, it does not account for the occurrence of zero markers. This is taken as evidence to support a differential non-development scenario of zero marking rather than a phonetic reduction scenario.

Keywords:
token-based
typology,
corpus typology,
zero marking,
zero exponence

INTRODUCTION

1

The present study examines the distribution of zero markers in adjectival, nominal, and verbal inflectional morphology.¹ In typology,

¹ I wish to thank the participants of the Freiburg Linguistics reading group for their helpful comments on earlier versions of this study: Uta Reinöhl, Peter

zero marking plays an important role for coding efficiency or form-frequency effects in morphosyntax. The analysis of form-frequency effects goes back to the early findings by Zipf (1935) that more frequent lexical elements tend to be shorter than less frequent ones. There is cross-linguistic evidence that, in inflectional morphology as well, more frequent or predictable markers tend to be shorter or at least not longer than comparable less frequent markers (Greenberg 1966; Guzmán Naranjo and Becker 2021; Haspelmath 2008b; Haspelmath *et al.* 2014; Haspelmath 2021; Haspelmath and Karjus 2017; Stave *et al.* 2021).

Such effects can be subsumed under the term of coding efficiency. The coding of grammatical expressions is efficient, because it saves effort in the production and processing of speech but maintains the successful transfer of information (cf. Levshina 2022, for an overview of efficiency in language and communication).

Usually, zero markers (in the sense of zero exponence) are grouped with shorter markers as opposed to longer ones. It is often explicitly or implicitly assumed that zero markers are used to express highly frequent morphosyntactic functions similarly to shorter markers (e.g. Bybee 2011; Croft 2003, Ch. 4; Diessel 2019, Ch. 11; Greenberg 1966, 32–37; Haspelmath 2008a, 2008b, 2021; Song 2018, Ch. 7). However, a quantitative cross-linguistic overview of the distribution of zero marking in inflection is still not available. The objective of this paper is to start filling this gap.

To do so, I analyze the distribution of zero markers in the UniMorph dataset (McCarthy *et al.* 2020), a cross-linguistic database of inflectional paradigms for individual lemmas. I first provide some theoretical background on zero marking and coding efficiency and introduce a working definition of zero markers in Section 2. Section 3 describes the dataset as well as the marker extraction procedure, and discusses examples of zero markers. I then analyze the probability of zero marking using the UniMorph dataset in Section 4. As will be seen, zero marking is generally dispreferred across languages and parts-of-

Arkadiev, Matías Guzmán Naranjo, Marvin Martiny, and Naomi Peck. I also thank the three anonymous reviewers for their valuable comments on earlier versions of this paper. This paper was supported by a Junior Fellowship from the Freiburg Institute for Advanced Studies, University of Freiburg.

speech. Section 5 then zooms in on those cells and values of adjectival, nominal, and verbal inflectional paradigms that are most likely to be zero marked across languages. In Section 6, I turn to the distribution of zero markers in language use. Using corpus data from the Universal Dependencies treebanks (Zeman *et al.* 2023), I analyze the association between token frequencies of inflection markers and their phonological length, including the distribution of zero markers. As we will see, frequency does not affect zero markers in the same way as it affects overt markers. Section 7 discusses the findings of this study with a special focus on the role of coding efficiency to account for the distribution of zero marking. Section 8 concludes.

ZERO MARKING

2

This section presents the relevant theoretical notions related to zero marking. Section 2.1 introduces zero marking and its relation to coding efficiency in typology. In Section 2.2, I propose a working definition of zero markers for the purposes of the present study. Throughout the paper, I use zero marking to refer to the absence of phonetic exponence (“zero exponence”) of a morphosyntactic function.

Zero marking and coding efficiency

2.1

The modern understanding of coding efficiency began with Zipf (1935), who showed that more frequent words tend to be shorter than less frequent words. Greenberg (1966, 1963) was one of the first typologists to relate the token frequencies of grammatical values to their formal markedness. An “unmarked” value in this sense is characterized by the absence of an exponent, which is contrasted with a “marked” value that is expressed by an overt exponent. For instance, Greenberg (1966, 32–37) showed how the markedness of singular, plural, and dual forms of nouns, verbs, and adjectives is reflected in their distribution in corpora from various languages. He noted that the formally unmarked (no exponent) number value, singular, is substantially more frequent than the formally marked number values (overt exponent) of plural and dual in corpus data from different languages.

Taking up Greenberg's findings but doing away with the concept of markedness, Haspelmath (2008a,b) argued that the length, complexity, and availability of grammatical markers can be accounted for by their frequency in language use. In a more recent study, Haspelmath proposed the following hypothesis:

- (1) *The grammatical form-frequency correspondence hypothesis*
When two grammatical construction types that differ minimally (i.e. that form a semantic opposition) occur with significantly different frequencies, the less frequent construction tends to be overtly coded (or coded with more segments), while the more frequent construction tends to be zero-coded (or coded with fewer segments), if the coding is asymmetric. (Haspelmath 2021, 2)

This hypothesis includes the assumption that zero forms pattern with shorter forms in being used to encode comparatively frequent expressions. Applied to inflectional morphology, we should thus expect zero marking for highly frequent values of morphosyntactic features. By now there is indeed much evidence for effects of coding efficiency between comparable grammatical expressions. However, examples usually only involve a difference in length, i.e. shorter vs. longer forms.² The participation of zero forms has not yet been the focus of any systematic cross-linguistic study. There are some indications from the literature, though, which suggests that coding efficiency and frequency may not be a suitable explanation for the distribution of zero markers. Stolz and Levkovych (2019) provide a qualitative overview of the distribution of zero marking in inflection ("absence of material exponence, AOME") from the perspective of canonical morphology. They note that "[f]rom the small number of cases discussed above it transpires that frequency might not always be the most powerful factor

² A few examples of quantitative approaches to form-frequency effects in grammar are: Guzmán Naranjo and Becker 2021 for the length and paradigmatic distribution of nominal inflection markers, Stave *et al.* 2021 for the length and frequency of morphemes in general, Haspelmath *et al.* 2014 for the expression of causal and non-causal alternations, Haspelmath 2008c for reflexive marking, Haspelmath and Karjus 2017 for number marking, and Ye 2020 for (in)dependent possessor marking.

to make a given word-form or category a candidate for AOME” (Stolz and Levkovych 2019, 396–397).

Guzmán Naranjo and Becker (2021) come to a similar conclusion based on a quantitative analysis of the association between the length of nominal inflection markers and their distribution across paradigms. They also use the UniMorph database, but focus on nominal inflection and test different distributional factors for their association with marker length. Although they find that marker length is associated with their type frequency, their results suggest that other measures such as the entropy of the marker are better predictors for their length. With their main focus being on predicting marker length from distributional measures, one detail of their analysis concerns zero marking and is highly relevant for the present study. Guzmán Naranjo and Becker (2021) note that a simple Poisson model to predict marker length strongly overestimates the occurrence of zero markers. This suggests that the distribution of zero markers does not simply follow the pattern of shorter ones.

Another area in which zero marking has been mentioned to behave differently is the occurrence of zero markers for person and number marking on verbs. Several quantitative typological studies (Bickel *et al.* 2015; Cysouw 2003; Siewierska 2010) find that zero marking for person marking is rather uncommon across languages. In contrast to the traditional view in typology, these studies do not find evidence for a paradigmatic preference of third person (singular) being zero marked on the verb. However, all three studies show that if a person marker is zero, it more likely expresses third person (singular) than first or second person.

Seržant and Moroz (2022) also mention zeros in verbal person-number marking. Analyzing the length of person-number markers in a typological sample, they argue for an attractor state in which the lengths of different indexes are associated with their frequencies in language use. Seržant and Moroz (2022, 6) note that “[...] articulatory efficiency plays an important role here: the more expected the sign is the shorter it is. Nevertheless, zero is not preferred.” They motivate the cross-linguistic avoidance of zero forms by invoking two types of efficiency: processing and planning efficiency. Seržant and Moroz (2022, 7) hypothesize that an overt exponent facilitates processing on the addressee’s side. They also propose that avoiding zero

marking makes planning more efficient on the speaker's side, "[...] because it provides a straightforward link from meaning to coding, while zero is inherently ambiguous by being linked to various meanings and domains" (Seržant and Moroz 2022, 7). Whether or not the avoidance of zero marking can indeed be accounted for by processing or planning efficiency requires proper psycholinguistic testing. The relevant point is that coding efficiency does not seem to be applicable to the frequency distribution of zero markers in person indexing in the same way as it is for overt markers.

2.2

A working definition of zero markers

The discussion and use of zero marking has a long tradition in morphology and in linguistics in general. It goes back to Pāṇini, who introduced the idea of zero morphs for morphemes that lack a phonetic representation as the outcome of morphological rules (Robins 1997, 181–182). The concept of zero morphs for linguistic analysis was also widely applied in later work by structuralists (e.g. Bloch 1947; Bloomfield 1933; Jakobson 1983[1939]; de Saussure 1916).³ Starting with Haas (1957), linguists began to criticize the assumption of zero morphs in the structuralist tradition and argued for stricter criteria to define zero morphs in order to avoid the assumption of excessive linguistic structure (e.g. Sanders 1988; Mel'čuk 2002; McGregor 2003). This was because linguists may postulate a zero morph for any single morphosyntactic function that does not correspond to an overt exponent. As Anderson (1992, 30) notes, it "leads to the formal problem of assigning a place in the structure (and linear order) to all of those zeros".⁴ Others, such as Arkadiev (2016), Contini-Morava (2006) and Mithun (1986), used data from typologically diverse languages to show that the absence of phonetic material can also correspond to the absence of a morphosyntactic feature rather than to

³For more details, see Meier 1961. See also Al-George 1967, Diehl 2008 and McGregor 2003 for more details on the history of linguistic zero.

⁴For examples and discussions of issues related to the use of zero morphs in morpheme-based, segmental approaches to morphology, see Anderson 1992, Pulum and Zwicky 1991, Blevins 2016 and Bank and Trommer 2015. For overviews of zero exponence in morphological theories, see Trommer 2012 and Dahl and Fábregas 2018.

zero marking. For instance, Lakota has overt markers for first and second person arguments on the verb, but no overt third person markers. Mithun (1986, 201–203) proposes to analyze the Lakota pattern as agreement that is restricted to first and second person arguments instead of analyzing agreement with third person arguments as zero marked.

In line with those more cautious approaches to zero morphs, this study uses the notion of “zero marker” as a descriptive shorthand for the absence of material exponence of a given morphosyntactic function (cf. Stolz and Levkovych 2019). In other words, I do not assume the presence of a zero morph. Instead, I understand zero markers as the absence of exponence which expresses a certain morphosyntactic function in addition to the lexical content of a word form. This also means that zero markers can only occur in contrast to at least one other, overtly coded morphosyntactic function of the same inflectional paradigm.

To analyze the distribution of zero markers in inflectional morphology, we need to identify the invariable, lexical parts (stems) as well as the potential exponents of a morphosyntactic function in an inflected word form. This conforms with the basic intuition that we want to separate the segments that convey the word’s lexical meaning from the segments that convey morphosyntactic information (cf. Matthews 1972).⁵ For the purposes of the present study, I define stems, markers, and zero markers as shown in (2), (3), and (4), respectively. These definitions are motivated by both theoretical and practical considerations regarding the dataset and annotations available.

(2) *Stem*

The stem expresses the lexical content of a word form; it corresponds to the longest common subsequence shared by all inflected forms of a word. The stem can be discontinuous.

(3) *Marker*

A marker encodes the morphosyntactic function of a word

⁵In reality, the identification of stems is not always this straightforward. There are many different ways in which the lexical parts of inflected words can vary in their phonological shape. Baerman and Corbett (2012) provide a number of examples and introduce a canonical approach to stems to capture that variation.

form, i.e. a value of some morphosyntactic feature defined for that word or a bundle of values of several such features. The marker corresponds to the phonetic material outside of the stem of a word form; it can be discontinuous.

(4) *Zero Marker*

A zero marker occurs when the word form does not feature any overt marker (as defined in (3)) to encode its morphosyntactic function. If the morphosyntactic function of the word consists of several morphosyntactic features, zero marking applies to the combination of feature values and not to single feature values in isolation.

Consider a simple example of stem and marker identification. The paradigm of English nouns consists of two cells: singular and plural. Given the paradigmatic relation between the singular form /dei/ (*day*.SG) and the plural form /dez/ (*day*.PL), we can identify the string /dei/ as the stem, i.e. the phonetic material that both forms of the paradigm share. Since the form filling the plural cell includes the additional material /z/, we can establish /z/ as a plural marker. In the singular cell, the form does not include any material other than what was identified as the stem. We can therefore treat the form of the singular cell of *day* as zero marked.

However, as will be described in detail in Section 3.3, I automatically adjusted the stems extracted according to the definition in (2) in order to account for stem allomorphy to a certain extent. This is motivated by the fact that many stem alternations are phonologically driven, which means that they do not necessarily provide meaningful insights about the inflectional properties of a system in general and about the distribution of zero marking in particular. Ignoring such alternations allocates additional material to the marker segments and runs the risk of systematically underestimating the number of zero markers. The adjusted marker_A and zero marker_A, which take into account stem alternations, are operationalized as described in (5) and (6), respectively.⁶

⁶From a theoretical perspective, it may be desirable to adjust the definition of stems and then derive the new definition of markers from that. The definitions given in (5) and (6) reflect the data extraction process, in that I extracted the ad-

- (5) *Marker_A*
A marker_A is extracted from a marker, as defined in (3), by removing all material from the affix positions that the system does not use for inflection.
- (6) *Zero Marker_A*
A zero marker occurs when the word form does not feature any overt marker (as defined in (5)) to encode its morphosyntactic function.

This operationalization of stems, markers_A and zero markers_A has the practical advantage that it does not require any morphological analysis particular to a single language or paradigm. It is a solution to identify the segments that contribute inflectional information that can be applied automatically and consistently to the cross-linguistic UniMorph dataset used in this study.

Besides practical considerations, this method is also based on theoretical grounds and follows the definition of stems by Beniamine and Guzmán Naranjo (2021), Bonami and Beniamine (2021), and Guzmán Naranjo and Becker (2021). Despite much theoretical work on the role and identification of stems in morphology, Bonami and Beniamine (2021) note that “there is no agreed upon method for identifying which part of an inflected word is a stem, and that the heuristics used by morphologists in that area are neither systematic nor principled enough”.⁷ They compare two types of stem identification based on prioritizing two different principles, namely to avoid stem allomorphy and to avoid discontinuous stems. Since those two principles are in conflict with each other many times, every approach to stem identification needs to rank them in some way to resolve such conflicts. Bonami and Beniamine (2021) compare the two methods of either adhering to the first or the second principle, resulting in what they call “unique discontinuous stems” (no stem allomorphy allowed) and “continuous stem sets” (no discontinuous stems allowed). While the

justed markers_A and zero markers_A from the original markers without extracting adjusted stems. I therefore omit the step of defining adjusted stems and focus directly on the alternative definitions of markers_A and zero markers_A.

⁷For work on stem identification and stem allomorphy, see Blevins 2003, Bonami 2012, Brown 1998, Maiden 1992, Montermini and Bonami 2013, Pirrelli and Battista 2000, Spencer 2012, Stump 2001 and Stump and Finkel 2013.

first method of unique discontinuous stems allocates all the variation of word forms to the exponents, leading to more exponent allomorphy, the second method of continuous stem sets keeps exponent allomorphy minimal, but leads to a high degree of stem allomorphy, since all variation that is enclosed by stem segments has to be included in the stems. What this shows is that neither approach creates more allomorphy; they simply allocate it differently. Of course, which of the two approaches is more useful depends on the research question at hand.

One of the questions discussed by the authors is what types of stems are more helpful in addressing the ‘Inflected Word Recognition Problem’ (IWRP), i.e. understanding what allows speakers to draw inferences from a word’s form about its content. This results in the task of separating the lexical and the inflectional parts of a word form, and Bonami and Beniamine (2021) note that “[i]n terms of the IWRP, the answer is quite simple. Sets of continuous stems are by definition less useful than a unique discontinuous stem: the unique discontinuous stem identifies exactly that part of the word that has no exponential value, while stem allomorphs blur the distinction between exponential and nonexponential material.” As the identification of zero forms relies on separating lexical segments from exponents of morphosyntactic information in word forms, the IWRP is of high relevance to this study and provides the theoretical grounds for the definition of stems given in (2).

This study will largely follow a word and paradigm approach to inflection (cf. Anderson 1992; Blevins 2016; Hockett 1967; Matthews 1972; Robins 1959; Stump 2001; Zwicky 1985). This approach bases morphological analyses on the paradigmatic relation between different word forms that represent the different morphosyntactic functions a given word can have. The exponent of a cell in an inflectional paradigm is determined through the relation of that word form to the forms used for the other cells of the paradigm. The word and paradigm approach has a very important practical advantage. It allows us to refrain from further segmentation of exponents into morphemes, which may require language-specific insights and which may not always be desirable or useful (cf. Blevins 2005, 2006).

Although morphological segmentation analyses may sometimes be uncontroversial, there are many cases where a morpheme analysis is less than clear. Various examples are given in Spencer 2012, one of

them being the Spanish subjunctive verb form *cantaríamos* ‘we would sing’. Several theoretical motivations exist to segment this word form into morphemes in five different ways: (i) *cant-a-r-í-a-mos*, (ii) *cantaríamos*, (iii) *cant-a-ría-mos*, (iv) *canta-r-í-a-mos* and (v) *cantar-í-amos* (Spencer 2012, 93). The fact that these profoundly varying morphological analyses are motivated in the literature suggests that such morpheme segmentations are always theoretically guided, whether explicitly or implicitly. It is likely that segmentation into morphemes in lesser-studied languages involves even more theoretical uncertainty, given that we may know much less about morphological structure and its diachrony than for languages like Spanish.

As will be shown in more detail in Sections 3.3 and 3.4, cells of paradigms are defined by (a combination of) values of morphosyntactic features. For instance, the inflectional paradigms of German nouns combine the morphosyntactic features of case and number. While nouns are inherently specified for gender, each word form in context is also specified for number and case, so that each cell of the paradigm corresponds to a number-case combination (e.g. dative plural).

For the purposes of this study, I do not distinguish between an exponent for plural number and one for dative case. Instead, I treat the material in addition to the stem in the dative plural cell as the marker of the dative-plural function. When no additional phonetic material is used, this cell is then analyzed as being zero marked (cf. Table 9). I do not assign zero markers to single abstract morphosyntactic values but to the relevant value combinations of the inflectional paradigms. The theoretical reason for this is that exponents of morphosyntactic functions are defined based on the relations between the forms of the different cells of the inflection paradigm, which combine these functions. This also reflects the morphological reality of many (if not most) languages, in that morphosyntactic functions are usually not marked in isolation but often occur in combination. As mentioned above, it is not always trivial to justify a segmental analysis. The practical reason is that there is still no language-independent and theory-independent way of segmenting distinct morphosyntactic exponents, and such segmentations are not (yet) automatable. Since automatic processing is indispensable for the purposes of the present study, no further segmentation of morphosyntactic exponents will be carried out.

The segmentation into stems and markers is often additionally complicated by inflection classes, which use different exponents to signal grammatical functions. Sections 3.3 and 3.4 show in more detail how the present approach deals with variation in the exponents due to inflection classes, with stem alternations and with suppletive forms.

3 DATASET AND SEGMENTATION

3.1 *Dataset*

The data used in this study comes from the UniMorph database (McCarthy *et al.* 2020), a large-scale cross-linguistic database of complete inflectional paradigms of adjectives, nouns, and verbs for individual lexemes from different languages. The present study includes adjectival, nominal, and verbal paradigms for 39, 62, and 96 languages, respectively. Some languages are featured with paradigms for more than one part-of-speech; a total of 114 languages is analyzed in this study. Figure 1 shows the geographical distribution of the languages in the dataset.⁸

While the dataset is not a balanced typological sample in the strict sense, it does include languages from all six macro areas (Africa, Eurasia, Papunesia, Australia, North America and South America), which ensures that typological and areal diversity is captured at least to some degree. Table 1 provides an overview of the final dataset with the number of languages, lemmas, paradigm cells, marker types and observations by part-of-speech. The morphosyntactic annotation in the UniMorph dataset follows the guidelines described by Sylak-Glassman (2016, 3), who notes: “This paper presents the Universal Morphological Feature Schema (UniMorph Schema), which is a set of morphological features that functions as an interlingua for inflectional morphology by defining the meaning it conveys in language-independent

⁸More details about the languages, the parts-of-speech, and the number of lexemes is provided in the files *affixation.csv* and *lemmas.csv* in the supplementary materials. All supplementary materials referred to in this paper can be found here: https://osf.io/p4mkc/?view_only=5238ace9cb1d4f4d998486ebb28f4fd8

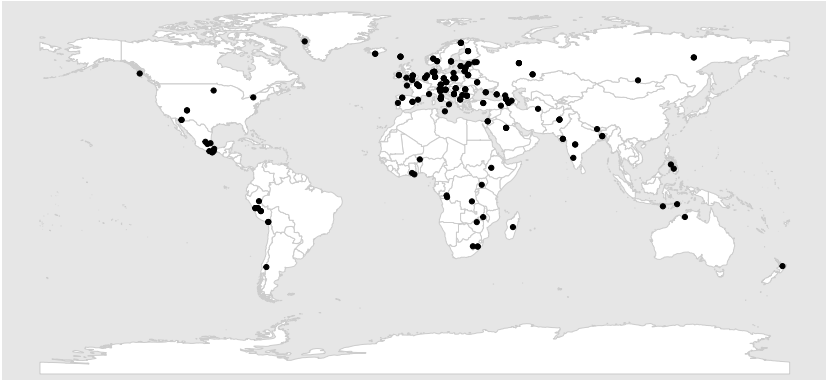


Figure 1:
Location
of the languages
in the dataset

	N langs	N lemmas	N cells	N markers	N obs
adjectives	39	157355	961	5552	6348198
nouns	62	610242	727	19537	6261881
verbs	96	129377	2753	47457	4407743

Table 1:
Dataset overview

terms. The features of the Universal Morphological Feature Schema have precise definitions based on attested cross-linguistic patterns and descriptively-oriented linguistic theory, and can capture the maximal level of semantic differentiation within each inflectional morphological category.” Annotations thus do not necessarily follow the linguistic traditions of particular languages but are defined and used in the sense of comparative concepts in typology (cf. Haspelmath 2018).

Data pre-processing

3.2

I excluded a number of languages available in UniMorph from the final analysis on the basis of unclear or insufficient annotations in the original datasets, some of which were annotated only automatically with no manual checks. Since the database is somewhat biased towards languages spoken in Eurasia (mostly Indo-European languages), I only included languages from this area with paradigms for more than 30 lemmas. For languages from other macro areas, especially from Africa or the Americas, I did not apply this threshold of 30 lemmas in order to

include more non-Indo-European languages and to keep the dataset as diverse as possible.⁹

The next step was to pre-process the data to remove errors and to make annotations more consistent across languages.¹⁰ Pre-processing consisted of different global and dataset-specific corrections. Global corrections included resolving annotation inconsistencies across languages. For example, the value “indefinite” was coded as “INDF” in some languages and as “NDEF” in others. Similarly, the annotation of person-number combinations in verbs varied, e.g. between “SG;1”, “1;SG”, “1SG” for first person singular. In such cases, I adjusted the annotation to a single label across all languages. I also removed complex lemmas containing a space or “-”. This removed some erroneous lemmas that were complex expressions rather than nouns, adjectives, or verbs. In some languages, both parts of a complex noun or adjective are inflected. Keeping such lemmas would have caused the marker extraction to detect infixation for complex lemmas with suffixes on two or more parts. Removing them avoided the artificial creation of more complex inflection patterns. Similarly, periphrastic forms were removed in the case of inflected auxiliaries, which would also have led to the erroneous analysis of infixation. This conservative approach of removing such forms was chosen over, e.g., splitting them or analyzing the inflected auxiliaries only. This alternative would have involved many additional case-by-case modifications of the original data, which in turn would have made it more prone to additional errors. Moreover, it would have increased the number of inflected forms from single auxiliaries, potentially misrepresenting the distribution of markers across lemmas. Complex forms were also removed if they contained a separate marker that occurred before or after the inflected verb form, depending on the cell of the paradigm. This was especially common with verbal paradigms, e.g.

⁹For adjectives, only Zulu has fewer than 30 lemmas (17); for nouns, this is the case only for Kalaallisut (23). For verbal paradigms, the languages with fewer than 30 lemmas are Sotho (26), Mapudungun (26), Murrinpatha (29), and Zarma (27).

¹⁰Detailed documentation of all pre-processing steps can be found in `preprocessing.txt` in the supplementary materials. For the implementation, see `code-preprocessing.R`.

verbal particles in German, or reflexive markers in Italian and Macedonian.

Dataset-specific cleaning steps included deleting “?” following interrogative verb forms in the Turkish data or deleting the indefinite article from Romanian nominal forms. Other cleaning steps were related to the alphabetic scripts used. For example, the Serbian-Bosnian-Croatian dataset contained forms in the Latin script with a handful of forms in Cyrillic. The latter were removed to allow for consistent processing. Some datasets (e.g. Old French or YoloXochitl Mixtec) contained alternative forms for certain cells. In such cases, I systematically kept the first form and removed the other(s).¹¹ Other dataset-specific operations included deleting single forms containing obvious errors (e.g. misalignment, cells with missing data).

Following data cleaning, I added phonological transcriptions to the inflected forms whenever possible. For some languages (e.g. Palantla Chinantec), the UniMorph database already provided the inflected forms in a phonological transcription. For most other languages, however, forms were given in the standard orthographic representation. This may well be problematic, especially for languages such as French, where the orthographic representation continues to make many distinctions that are no longer realized in the spoken language. For this reason, whenever possible, I replaced the orthographic forms by a phonological transcription using Epitran (Mortensen *et al.* 2018). Epitran currently has modules to transcribe 31 of the languages used here.¹²

While not perfect, Epitran offers a more realistic representation of the forms occupying the different cells of inflectional paradigms. Table 2 illustrates this by showing the transcriptions generated with

¹¹ It would have been insightful to include overabundance in a systematic way. Overabundance refers to the phenomenon of two distinct forms being available to express a single cell in a paradigm (cf. Thornton 2012). However, alternative forms are not systematically annotated in the UniMorph datasets. If provided, their relation differs greatly across datasets and is not usually documented in the dataset descriptions. Alternatives can represent diachronic, dialectal, or stylistic variants; in other cases, their alternation behavior remains unclear. It is also unclear how many overabundant forms are not provided in UniMorph. Including overabundance is thus not possible with the approach used in this study.

¹² For details, see `epitran.py` in the supplementary materials.

Table 2:
Phonological
transcription of
the French verb
allumer ‘turn on
(light)’

	1SG	2SG	3SG	1PL	...
PRS.IND	<i>allume</i> alym	<i>allumes</i> alym	<i>allume</i> alym	<i>allumons</i> alymiõ	
PST.IPFV.IND	<i>allumais</i> alyme	<i>allumais</i> alyme	<i>allumait</i> alyme	<i>allumions</i> alymiõ	
PST.PFV.IND	<i>allumai</i> alyme	<i>allumas</i> alyma	<i>allumat</i> alyma	<i>allumâmes</i> alymam	
FUT	<i>allumerai</i> alymre	<i>allumeras</i> alymra	<i>allumera</i> alymra	<i>allumerons</i> alymreõ	
PRS.COND	<i>allumerais</i> alymre	<i>allumerais</i> alymre	<i>allumerait</i> alymre	<i>allumerions</i> alymriõ	
PRS.SBJV	<i>allume</i> alym	<i>allumes</i> alym	<i>allume</i> alym	<i>allumions</i> alymiõ	
PST.SBJV	<i>allumasse</i> alymas	<i>allumasses</i> alymas	<i>allumât</i> alyma	<i>allumassions</i> alymasiõ	
...					

Epitran for the French verb *allumer* ‘light something, turn on (light)’. The rows show seven TAM combinations; for each of these, the first row contains the form in orthographic representation, and the second row shows the phonological transcription generated with Epitran. For the remaining 81 languages, the forms in UniMorph are given in their orthographic representation, which reflect the phonological shapes to a varying degree. To consider the potential influence that the type of phonological representation may have on the detection of zero forms, I manually coded whether or not the representation was phonological.¹³ Orthographic representations that systematically reflected phonology were treated as phonological representations. This led to 31 languages with a transcription generated using Epitran, 63 languages with original representations that systematically reflect phonological shape, and 20 languages with orthographies that do not always reflect phonological shape. The type of phonological representation was then added as a control variable in the analysis.

¹³For details by language, see *affixation.csv* in the supplementary files.

In order to analyze the distribution of zero markers, I automatically segmented the inflected word forms following the method developed in Beniamine and Guzmán Naranjo 2021 and Guzmán Naranjo and Becker 2021. As mentioned in Section 2.2, the segmentation follows a word and paradigm approach to morphology, in that whole forms are paired with morphosyntactic functions according to their distribution across the inflectional paradigms. This means that the subsequence shared by all cells of the paradigm is automatically extracted and taken as the stem according the working definition given in (2). All material not included in this subsequence is analyzed as the marker of a given cell, as defined in (3). If the form corresponds to the longest common subsequence (i.e. the stem), the marker is analyzed as zero according to the definition in (4). This automated detection of stems and markers is necessary for two reasons. First, it is not feasible to apply manual, language-specific segmentations to this dataset. Second, this method allows for a single, consistent way of detecting zero marking across languages, which is necessary for the cross-linguistic comparisons made in this study.¹⁴

To give a simple example of the segmentation into stems and markers and of the detection of zero markers, Table 3 shows parts of the present tense paradigm of the French verb *allumer* from Table 2.¹⁵ Comparing the forms of the different cells of the paradigm, the string *alym* is identified as the longest common subsequence between all forms of the paradigm. For the purposes of the present paper, this subsequence is analyzed as the stem. All remaining material is analyzed as the marker of a particular cell. In cells where the form corresponds to the stem, markers are analyzed as zero. This is the case for some of the present tense forms; such cells are shaded in grey in Table 3.

In the remainder of this section, I discuss the extraction of stems and markers using examples that may appear less straightforward, in

¹⁴Stem alternations are not accounted for by this extraction method; Section 3.4 shows how they are included in the present study.

¹⁵This example involves a continuous stem as well as continuous markers. Examples of discontinuous stems are shown later in this section and in Section 3.4.

Table 3:
Marker extraction for the French verb
allumer ‘turn on (light)’

Cell	Form	Stem	Marker
PRS.IND.1SG	alym	alym	-
PRS.IND.2SG	alym	alym	-
PRS.IND.3SG	alym	alym	-
PRS.IND.1PL	alymən	alym	-ən
PRS.COND.1SG	alymere	alym	-ere
PRS.COND.2SG	alymere	alym	-ere
PRS.COND.3SG	alymere	alym	-ere
PRS.COND.1PL	alymerjən	alym	-erjən
PRS.SBJV.1SG	alym	alym	-
PRS.SBJV.2SG	alym	alym	-
PRS.SBJV.3SG	alym	alym	-
PRS.SBJV.1PL	alymjən	alym	-jən
...

that the identified stems (and thus also markers) do not correspond to stems as traditionally analyzed in the literature, or in that they are discontinuous.

One example comes from Ayamara (Aymaran), a language with nominal inflection known for its subtractive morphology. The accusative singular cell is usually analyzed as being expressed by the subtraction of the final vowel of the nominative singular form (cf. Coler 2015). Table 4 illustrates this with parts of the paradigms of two Aymara nouns. For the purposes of this study, the accusative singular form corresponds to the stem, because it equals the longest common subsequence of all forms of the lexeme. Compared to the accusative form, the nominative form has an additional final vowel, which is also found in all other forms of the paradigm, except for the inessive (INESS) and equative (EQTV) forms.

Traditionally, the nominative form with the final vowel is analyzed as the stem of the noun, while the accusative is argued to be a subtractive form, i.e. consisting of less material than the stem of the lexeme (Baerman *et al.* 2017; Coler 2015, 2018). There are valid diachronic arguments to support such an analysis. Coler (2018) provides examples of historical Aymara with accusative forms that still have the final vowel. In addition, vowel deletion is a common phonological

Cell	Form	Stem	Marker	Form	Stem	Marker
NOM.SG	anu	an	-u	chaski	chask	-i
ACC.SG	an	an	-	chask	chask	-
GEN.SG	anuna	an	-una	chaskina	chask	-ina
COM.SG	anumpi	an	-umpi	chaskimpi	chask	-impi
ABL.SG	anuta	an	-uta	chaskita	chask	-ita
ALL.SG	anuru	an	-uru	chaskiru	chask	-iru
INESS.SG	anpacha	an	-pacha	chaskpacha	chask	-pacha
EQTV.SG	anjama	an	-jama	chaskjama	chask	-jama
...

Table 4:
Marker
extraction
for the Aymara
nouns *anu* ‘dog’
and *chaski*
‘messenger’

process in Aymara. Nevertheless, aiming at a synchronic, comparable analysis across languages here, I treat the accusative form as the stem of the lexeme. In the Aymara data, the accusative is zero marked in all 1,522 nouns of the dataset, without exception.

Another rather unusual case of zero marking can be found in Georgian (Kartvelian) verbs. Besides a number of other theoretically interesting patterns, Georgian verbs have been cited in the typological and morphological literature for their cross-linguistically unusual 2SG zero marker (e.g. Anderson 1992; Blevins 2016; Stolz and Levkovych 2019). However, not all lexemes express the 2SG form with a zero marker in the sense of the present study. Only one out of 118 verbal lexemes in the dataset features a zero marker in the 2SG present tense cell. Table 5 shows this for the verb *ts’ers* ‘write’, in opposition to *ak’eteb* ‘make’.¹⁶

In general, Georgian verbs take a so-called preverb in some but not all of the tenses (Hewitt 1995, 148–169). When present, it precedes the prefixal part of agreement marking on the verb. As we can see in Table 5, present and imperfect forms occur without the verbal prefix, while the future, aorist, and perfect forms all make use of the prefix (*da-* and *ga-* in the examples in Table 5). In most TAM series, many Georgian verbs also have so-called thematic suffixes (Hewitt 1995, 143–147), such as *-eb* in *ak’eteb* ‘make’. The presence of such thematic suffixes in the present tense results in the absence of

¹⁶ The segment *-a-* is not part of the verb stem of *ak’eteb* ‘make’, as it does not occur in all forms of the paradigm, e.g. the imperfective masdar form *k’etebi*.

Table 5:
Marker
extraction
for the Georgian
verbs
ts'ers 'write'
and *ak'eteb*
'make'

Cell	Form	Stem	Marker	Form	Stem	Marker
PRS.1SG	vts'er	ts'er	v-	vak'eteb	k'et	va-eb
PRS.2SG	ts'er	ts'er	-	ak'eteb	k'et	a-eb
PRS.1PL	vts'ert	ts'er	v-t	vak'etebt	k'et	va-ebt
IMPF.1SG	vts'erde	ts'er	v-de	vak'etebdi	k'et	va-ebdi
IMPF.2SG	ts'erde	ts'er	-de	ak'etebdi	k'et	a-ebdi
IMPF.1PL	vts'erdet	ts'er	v-det	vak'etebdit	k'et	va-ebdit
FUT.1SG	davts'er	ts'er	dav-	gavak'eteb	k'et	gava-eb
FUT.2SG	dats'er	ts'er	da-	gaak'eteb	k'et	gaa-eb
FUT.1PL	davts'ert	ts'er	dav-t	gavak'etebt	k'et	gava-ebt
AOR.1SG	davts'ere	ts'er	dav-e	gavak'ete	k'et	gava-e
AOR.2SG	dats'ere	ts'er	da-e	gaak'ete	k'et	gaa-e
AOR.1PL	davts'eret	ts'er	dav-et	gavak'etet	k'et	gava-et
...

zero marking for most of the verbs. The thematic suffix *-eb/-ob* is part of the second person singular present form; as it is not used in the aorist forms, the former does not correspond to the longest common subsequence of the verb forms. The second person singular present-tense cell can thus only be expressed by a zero form with verbs that generally do not use any of the thematic suffixes, like the verb *ts'ers* 'write' in Table 5.

Arabic (Semitic) is well known to have roots that consist of discontinuous consonants, with prefixed, infixes, and suffixed vowels, and other consonants to mark the grammatical values of a given form in the paradigm (e.g. Boudelaa and Marslen-Wilson 2001; Ratcliffe 1998; Schramm 1962; Yip 1988). The automatic extraction of the longest common subsequence detects these consonants and assigns all additional material to the markers. This is shown for two verbs, *?arsala* 'send' and *iktašafa* 'discover' in Table 6.

Another language that is interesting from the point of view of marker extraction is Tohono O'odham (Uto-Aztecan, Mexico, USA). Some nouns in Tohono O'odham mark the plural using partial reduplication of the stem (Hill and Zepeda 1998). Table 7 shows this for the two nouns *ban* 'coyote' and *ceoj* 'boy', using the phonological transcription generated by Epitran.

Cell	Form	Stem	Marker	Form	Stem	Marker
IPFV.1SG	ʔursilu	rsl	ʔu-i-u	ʔaktašifu	ktšf	ʔa-a-i-u
IPFV.2SG.F	tursilina	rsl	tu-i-ina	taktašifina	ktšf	ta-a-i-ina
IPFV.3PL.M	yursilūna	rsl	yu-i-ūna	yaktašifūna	ktšf	ya-a-i-ūna
PFV.1SG	ʔarsaltu	rsl	ʔa-a-tu	iktašaftu	ktšf	i-a-a-tu
PFV.2SG.F	ʔarsalti	rsl	ʔa-a-ti	iktašafti	ktšf	i-a-a-ti
PFV.3PL.M	ʔarsalū	rsl	ʔa-a-ū	iktašafū	ktšf	i-a-a-ū
...

Table 6:
Marker
extraction
for Arabic verbs
ʔarsala ‘send’
and *iktašafa*
‘discover’

Cell	Form	Stem	Marker	Form	Stem	Marker
SG	ban	ban	-	$\overline{\text{tjind}\overline{\text{z}}}$	$\overline{\text{tjind}\overline{\text{z}}}$	-
PL	ba:ban	ban	:-ba-	$\overline{\text{tjitjind}\overline{\text{z}}}$	$\overline{\text{tjind}\overline{\text{z}}}$	-tj-

Table 7:
Tohono O’odham
nouns *ban*
‘coyote’ and *ceoj*
‘boy’

Applying the automatic stem extraction for the purposes of this study, the reduplicated stem is analyzed as infixation, i.e. the marker of the plural cell occurs within the sequence shared by both cells.

Stem alternations and suppletion

3.4

The previous examples showed that stems correspond to continuous strings to a differing degree; in fact, alternations within stems are common across languages. Stem alternations can be defined as phonological changes within the material expressing the lexical meaning of a word across the cells of a paradigm (cf. Paster 2016; Baerman and Corbett 2012). As was mentioned in Section 2.2, such alternations do not necessarily provide meaningful insights about the inflectional properties of a system. For inflected forms with stem alternations, the stem and marker extraction method shown in Section 3.3 would result in material being analyzed as part of the marker that could otherwise be considered as belonging to the stem. Therefore, this method runs the risk of detecting fewer zero markers than potentially there could be.

To gauge the effect of marker material resulting from stem alternations, I extracted another set of zero markers_A, as defined in (5), by removing material that could be analyzed as a stem alternation. To do so, I determined the position(s) of inflectional affixation for all language and part-of-speech combinations in the dataset. This was done

Table 8:
Marker_A
extraction

Affix position	Removal	Marker	Marker _A
pfx	remove infixes and suffixes	pfx-ix-sfx	pfx-
sfx	remove prefixes and infixes	pfx-ix-sfx	-sfx
pfx + sfx	remove infixes	pfx-ix-sfx	pfx-sfx
ix + sfx	remove prefixes	pfx-ix-sfx	-ix-sfx
pfx + ix + sfx	/	pfx-ix-sfx	pfx-ix-sfx

based on language descriptions and on the extracted stems and markers used in this study. Given the observed patterns, I distinguished between the following five categories of affix position: prefix, suffix, prefix + suffix, infix + suffix, prefix + infix + suffix.¹⁷ Using this classification, all material that had originally been assigned to the marker but did not occur in a regular affix position for a given language and part-of-speech was removed. A schematic overview of this step is shown in Table 8. For instance, if a language and part-of-speech combination is classified as having prefixes only, all additional material that would be classified as an infix or suffix was removed. Similarly to the first step of stem and zero marker extraction, these marker adjustments were automated so that they could be applied systematically for all the languages in the dataset without any additional manual annotations. For the type prefix + infix + suffix only, no additional material could be removed from markers, because all available affix positions were already used by inflectional morphology. The three languages in this category are Arabic, Hebrew, and Maltese; I applied no further changes to the markers in these cases.

The following paragraphs provide a few examples of how markers_A, as defined in (5) and (6) (cf. Section 2.2), were extracted in the presence of stem alternations. One example is a vowel change in German nouns, where a back stem vowel in the singular cells is opposed to a front stem vowel in the plural cells. This is shown for the German noun *Kloß* ‘dumpling’ in Table 9. All forms are given in the phonological transcription generated with Epitran.

In the case of *Kloß*, the longest common subsequence is not continuous. Due to the umlaut in the plural forms, the automatically ex-

¹⁷The list of languages and affix position values can be found in `affixation.csv` in the supplementary materials.

Cell	Form	Stem	Marker	Marker _A
NOM.SG	klos	kls	-o-	-
ACC.SG	klos	kls	-o-	-
DAT.SG	klos	kls	-o-	-
GEN.SG	kloses	kls	-o-es	-es
NOM.PL	kløsa	kls	-ø-a	-a
ACC.PL	kløsa	kls	-ø-a	-a
DAT.PL	kløsa	kls	-ø-a	-a
GEN.PL	kløsa	kls	-ø-a	-a

Table 9:
Marker extraction
of the German noun
Kloß ‘dumpling’

tracted stem of *Kloß* consists of the three consonants *kls*. The vowel change from /o/ in the singular to /ø/ in the plural is analyzed as a part of the cells’ markers, respectively. Therefore lemmas such as *Kloß* in German do not have zero marking according to the first method of marker extraction. However, German nouns are classified as using suffixes only for inflection. Adjusting the markers by removing all material that is not a suffix takes into account that the alternation between /o/ and /ø/ is a stem alternation. The markers_A no longer contain infixal material and are analyzed as zero for the nominative, accusative, and dative singular cells. Another process of stem alternation is metathesis. Table 10 shows how this is dealt with in the case of the Hungarian noun *gyomor* ‘stomach’. In this example, the final segment *-or* is metathesized when certain affixes are added to the stem. Again, this leads to a situation where the stem does not include the segment undergoing metathesis, and the discontinuous string *jomr*

Cell	Form	Stem	Marker	Marker _A
NOM.SG	jomor	jomr	-o-	-
ACC.SG	jomrot	jomr	-ot	-ot
DAT.SG	jomornøk	jomr	-o-nøk	-nøk
INSTR.SG	jomor:ðl	jomr	-o:ðl	-:ðl
TERM.SG	jomorig	jomr	-o-ig	-ig
ON.ESS.SG	jomron	jomr	-on	-on
ON.ALL.SG	jomor:ð	jomr	-o:ð	-:ð
ON.ABL.SG	jomor:o:l	jomr	-o:o:l	-:o:l
...

Table 10:
Marker extraction
for the Hungarian noun
gyomor ‘stomach’

Table 11:
Marker
extraction
for the Slovenian
adjective
absúrden ‘absurd’

Cell	Form	Stem	Marker	Marker _A
NOM.SG.M.INDF	absúrden	absúrdn	-e-	-
NOM.SG.N	absúrdno	absúrdn	-o	-o
NOM.SG.F	absúrdna	absúrdn	-a	-a
DAT.SG.M	absúrdnemu	absúrdn	-emu	-emu
DAT.SG.N	absúrdnemu	absúrdn	-emu	-emu
DAT.SG.F	absúrdni	absúrdn	-i	-i
...

is analyzed as the stem. This in turn leads to the infixal marker *-o-* in the NOM.SG cell, for instance. However, Hungarian only uses suffixation for nominal inflection, and the NOM.SG is usually (81% in this dataset) not overtly marked in Hungarian. Therefore the adjusted markers_A no longer feature material that is infixal, and the NOM.SG is zero marked for the noun *gyomor* as well.

Another example of stem-internal alternations is epenthesis, the addition of phonological material in the stem in some but not all of the cells in the paradigm. One example of epenthesis is found with certain types of adjectives in Slovenian, which feature stem-final consonant clusters. This can be seen with the adjective *absúrden* ‘absurd’ in Table 11. Similarly to the previous examples, Slovenian adjectives only use suffixation to mark inflection. In Table 11, in all but one inflected form, the stem ends in the cluster */rdn/*, and an overt suffix is added to the stem. The NOM.SG.M.INDF cell, however, is not marked by an additional suffix. Instead, the epenthetic vowel */-e-/* is inserted between the stem-final consonants to break up the consonant cluster. The adjusted markers_A remove all infixal material for Slovenian adjective markers, which results in the NOM.SG.M.INDF cell being analyzed as zero marked.

Stem alternations are relevant in yet another way in Tlapezco Chinantec (Otomanguean). This language has a complex inflectional paradigm, combining various patterns of stem and tone changes. Table 12 shows the inflectional paradigm of the verb *køgʔ²* ‘eat’. The forms of *køgʔ²* have different tones for first vs. second and third person forms in all three tenses. Given that the tones are represented by superscript numbers following the tone-bearing unit, they are taken into account by the extraction and detection of zero markers. While

Cell	Form	Stem	Marker	Marker _A
PRS.1SG	køɣʔ ¹²	køɣʔ	- ¹²	-
PRS.1PL	køɣʔ ¹²	køɣʔ	- ¹²	-
PRS.2	køɣʔ ²	køɣʔ	- ²	-
PRS.3	køɣʔ ²	køɣʔ	- ²	-
PST.1SG	mi ³ -køɣʔ ¹²	køɣʔ	mi ³⁻¹²	mi ³ -
PST.1PL	mi ³ -køɣʔ ¹²	køɣʔ	mi ³⁻¹²	mi ³ -
PST.2	mi ³ -køɣʔ ²	køɣʔ	mi ³⁻²	mi ³ -
PST.3	mi ³ -køɣʔ ²	køɣʔ	mi ³⁻²	mi ³ -
FUT.1SG	køɣʔ ¹³	køɣʔ	- ¹³	-
FUT.1PL	køɣʔ ¹³	køɣʔ	- ¹³	-
FUT.2	køɣʔ ³	køɣʔ	- ³	-
FUT.3	køɣʔ ¹	køɣʔ	- ¹	-

Table 12:
Marker extraction
for the Tlatepuzco
Chinantec verb *køɣʔ²* ‘eat’

present and future tense forms do not make use of an additional segmental marker, the tone annotations are extracted as marker material. Given that otherwise Tlatepuzco Chinantec verbs only use prefixation, I removed all infixal and suffixal material for the adjusted markers_A. As can be seen in Table 12, the adjusted markers_A now capture tonal changes as changes to the stem, and the present and future tense cells are now taken to be zero marked. Although this automated way of accounting for stem alternations is able to deal with almost all of the relevant cases, there is one type of alternation that this method cannot capture. If a stem alternation occurs at the edge between stem and affix, then the extraction methods used for this study are not able to detect that the boundary between marker and stem should occur in a different position.

One example is the so-called consonant gradation in Northern Saami (Uralic). It can be described as an alternation of the final stem consonants across the cells of the paradigm, leading to their weakening or strengthening (cf. Bakró-Nagy 2022). An example of Northern Saami adjectives is shown in Table 13. We see that the final stem consonant of the adjective *aiddolaš* ‘exact’ alternates between /-š/, /-čč/ and /-žž/. The extraction process used here analyzes this alternation as part of the marker. By contrast, the adjective *bahá* ‘angry’ shows the marker extraction for adjectives with no stem alternations. For such adjectives, the NOM.SG, ACC.SG, and GEN.SG cells are zero marked.

Table 13:
Marker
extraction for the
Northern Saami
adjectives
aiddolaš ‘exact’
and *bahá* ‘angry’

Cell	Form	Stem	Marker _(A)	Form	Stem	Marker _(A)
NOM.SG	aiddolaš	aiddola	-š	bahá	bahá	-
ACC.SG	aiddolačča	aiddola	-čča	bahá	bahá	-
GEN.SG	aiddolačča	aiddola	-čča	bahá	bahá	-
ILL.SG	aiddolažžii	aiddola	-žžii	bahái	bahá	-i
COM.SG	aiddolaččain	aiddola	-ččain	baháin	bahá	-in
FRML.SG	aiddolažžan	aiddola	-žžan	bahán	bahá	-n
PRP.SG	aiddolaččas	aiddola	-ččas	bahás	bahá	-s

Thus, in cases of alternation at the edge between the stem and the inflectional affix, this method of marker extraction is unable to detect zero marking.

In its most extreme form, a stem alternation that includes the edge segments of stems is suppletion. Suppletion refers to stem alternations where maximally different phonological forms are used to express the same lexical component of an inflected word form across different cells of the paradigm (cf. Mel’čuk 1994; Corbett 2007). Suppletive forms go beyond alternations that can be described in terms of phonological or prosodic relations between forms (at least synchronically). Consider the English examples given in Table 14, where we see the verbs *think* and *go*, both with suppletive stems. In the case of *think*, the suppletion does not affect the entire stem, as the initial segment *θ-* is found in all cells of the paradigm. As a consequence, the extracted marker ends up with all the remaining material (which would usually be analyzed as being part of a suppletive stem). In the case of *go*, suppletion is complete in that no segment is shared between all cells of the paradigm. The complete phonological strings of each form are thus extracted as markers of their respective cells. As the examples from Northern Saami and English showed, neither marker extraction method used

Table 14:
Marker
extraction for
the English verbs
think and *go*

Cell	Form	Stem	Marker _(A)	Form	Stem	Marker _(A)
NFIN	θɪŋk	θ	-ɪŋk	gow	-	gow
PRS.3SG	θɪŋks	θ	-ɪŋks	gowz	-	gowz
PTCP.PRS	θɪŋkɪŋ	θ	-ɪŋkɪŋ	gowɪŋ	-	gowɪŋ
PST	θɔt	θ	-ɔt	went	-	went
PTCP.PST	θɔt	θ	-ɔt	gɔn	-	gɔn

for this study has a principled way of removing alternating stem segments that are adjacent to affixal material from the marker. Therefore neither method detects potential zero marking with suppletive forms, as they will always assign phonological material to the marker. While it is possible to exclude markers that occur only once per cell (cf. Section 3.5), many suppletive forms do not correspond to such hapax legomenon markers. Especially larger datasets often include complex lemmas such as *overthink* or *undergo* in English, for example. The extracted markers *-gow* and *-ɲk* from Table 14 occur 11 times in the verbal paradigms of English. The stem alternation pattern shown for Northern Saami in Table 13 occurs systematically (26 times) in the dataset. In such cases, I do not have any principled way of excluding the markers from the analysis.

To remain agnostic about the effect of stem alternations and to apply a systematic approach to all languages, I performed the analyses in Sections 4 and 5 for both sets, markers and markers_A. Since the results were very similar with no substantial differences, I only report the results of markers_A, for reasons of brevity. Details about the results based on the originally extracted markers can be found in the supplementary materials as indicated in the relevant sections. Given that no substantial differences were found for the distribution of zero markers in inflection paradigms, I only analyze the distribution of markers_A in the corpus data in Section 6. Whenever markers are mentioned in the following sections, I refer to markers_A, if not stated otherwise.

Hapax legomenon markers

3.5

The dataset includes a number of markers that occur only once per cell for a given language and part-of-speech combination. Some of these hapax legomenon markers are the result of stem alternations, but most of them result from the remaining errors in the dataset. In total, I identified the following number of hapax legomenon markers: 9,223 for adjectives, 23,539 for nouns, and 54,768 for verbs. In terms of marker types, hapax legomenon markers make up a large proportion, namely 0.45, 0.46, and 0.42 for adjectives, nouns, and verbs, respectively. In terms of the total number of occurrences, however, they only amount to a proportion of 0.003 for adjectives, 0.008 for nouns, and 0.03 for verbs.

One example of a hapax legomenon marker as the result of stem alternation comes from Northern Saami. The adjective *čáppat* ‘pretty’ features gradation similarly to the example shown in Table 13. In this case, stem-final *-pp* alternates with *-bb* across cells of the paradigm. This type of alternation is only attested once in the dataset, making all markers extracted from the lemma *čáppat* hapax legomenon markers.

Most hapax legomenon markers, however, result from remaining material that is not part of the inflected word forms or from errors in the automatic phonological transcription performed by EpiTran. To give one example, in the Hungarian dataset, the impersonal verb *fái* ‘hurt’ features the string “only3rdpersonforms” as the verb form in a number of cells. This string is of course not a Hungarian verb form, but an additional linguistic annotation, which causes the extraction of the longest common substring to find nonsensical strings and hence hapax legomenon markers.

Visual inspection of the hapax legomenon markers suggests that most result from the automatic phonological transcription using EpiTran. For instance, the German adjective *makaberə* ‘macabre’ shows an alternation between stem-final *-b* and *-p* in the phonological transcription. All forms except the comparative form have *-b*, while the comparative form *makapɐv* has *-p*, which leads to hapax legomenon markers.

In order to exclude such markers, as they do not provide much insight into the distribution of zero marking, I removed all hapax legomenon markers from the dataset. Given that their proportion of the total number of observations is very low, it is safe to assume that their removal will not artificially distort the distribution of zero markers.

3.6

Morphomic paradigms

Another potential factor influencing the distribution of zero marking is the distribution of inflected word forms across the paradigm. Many paradigms have syncretic cells, where a single form expresses more than one cell. Taking this into account and considering only the different forms that are found in a paradigm may thus lead to different probabilities of zero markers. To examine how much the results change if

proportions of zero marking are established using distinct forms only, I collapsed the data into morphomic paradigms (cf. Boyé and Schalchi 2016). Morphomic paradigms consist of all the different forms that a given word can have without taking into account their meaning. Syncretic forms are counted in only once in morphomic paradigms. Section 4 therefore analyzes the distributions of markers in morphomic paradigms in addition to paradigms that include information on cells. The analysis of the effect of token frequency in language use on the distribution of zero marking in Section 6 is also based exclusively on forms, i.e. morphomic paradigms.

ESTIMATING THE PROBABILITY OF
ZERO MARKERS

4

Observed distributions

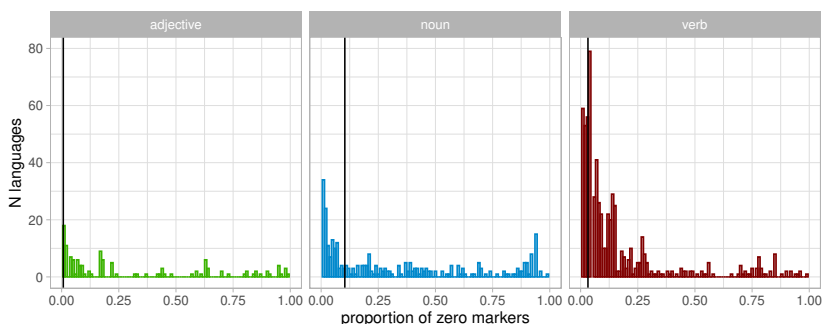
4.1

In order to examine the probability of zero markers in adjectival, nominal, and verbal inflection, Table 15 and Figure 2 provide an overview of the observed distribution of zero marking in inflection. The second column of Table 15, “N forms zero”, shows the number of inflected word forms across parts-of-speech that are zero marked. The third column, “prop forms zero”, indicates the proportion of zero-marked word forms in the entire dataset. We see that the proportions of zero markers are very low for adjectives; verbs show a somewhat higher proportion, and nouns have the highest proportions of zero marking at about 0.1. Zero marking is clearly not common in inflection of any of the parts-of-speech. The last two columns of Table 15 show the number of cells where zero marking is absent and the number where zero marking is used for all lemmas. Unsurprisingly, we find a high number of cells

pos	N forms	prop forms	N cells	
	zero	zero	no zero marking	all zero marking
adj	45,859	0.007	1,439	12
noun	648,859	0.104	1,227	5
verb	141,268	0.032	3,771	26

Table 15:
Observed
proportions
of zero markers

Figure 2:
Number
of languages
by part-of-speech
and proportion
of zero markers
(solid lines
correspond
to the overall
proportions
of zero markers)



with no zero marking at all, and only a very small number of cells that feature zero marking consistently across all lemmas.¹⁸ For the last two columns, we find an increasing number of cells from nouns to adjectives to verbs. This reflects the number of cells that those three parts-of-speech distinguish in the dataset, with 727, 961, and 2,753 cells for nouns, adjectives, and verbs, respectively. Figure 2 shows a histogram of the proportions of zero marking in adjectival, nominal, and verbal inflection. The overall proportions are indicated by a vertical line. We can see that they vary to a great extent across languages and parts-of-speech. All three parts-of-speech exhibit a preference for proportions of 0 or close to 0. This preference is most pronounced for adjectives and verbs. For nouns, we find a more balanced distribution, with more proportions above 0.5 for zero marking.

There are five additional factors that are relevant for estimating the probability of zero markers in inflection: the number of cells in a paradigm, the number of morphosyntactic values expressed per cell, the number of lemmas for which paradigms are available, the usual affix position, and the type of phonological representation. The number of cells in a paradigm can be taken as a measure of paradigm size. It is an important factor to include, since it is possible that zero markers are less likely to occur in a larger paradigm that makes more morphosyntactic distinctions. Table 16 gives an overview of the number of cells per paradigm in the dataset, showing the minimum, maximum,

¹⁸The figure of 26 cells that are expressed by zero markers exclusively is rather high; this can in part be explained by many cells in the verbal paradigm that only occur in single languages.

	min	max	Q1	median	Q3
adjective	3	256	13.5	26	51
noun	2	256	8.5	14	23.5
verb	2	432	15	30	50.5

Table 16:
Number of cells

median, the first and the third quartile. As the number of cells spans several magnitudes, I use log-transformed values for the analysis.

Another important factor for estimating the probability of zero marking is the number of morphosyntactic values expressed per cell.¹⁹ For the purposes of this study, we can take the number of values per cell to represent the semantic complexity of the inflectional markers. A summary of the number of values per cell is shown in Table 17.

	min	max	Q1	median	Q3
adjective	1	5	2	3	3
noun	1	4	2	2	2
verb	1	7	1.75	2	2.25

Table 17:
Number of morphosyntactic values per cell

Including this factor in the analysis is important, since one could expect more complex markers (which express more complex meanings) to be encoded by more material. The average number of lemmas for which inflectional paradigms are available is not inherently related to the probability of zero marking, but may influence it. As can be seen in Table 18, the median number of lemmas differs greatly across

	min	max	Q1	median	Q3
adjective	17	98464	131	507	1994
noun	23	235294	248	1240	4591
verb	26	30032	109	374	910

Table 18:
Number of lemmas

languages. It is therefore an important factor to be controlled for. Another factor that is included in the analysis for its potential effect on the probability of zero marking is the position of the marker regarding the stem. As described in Section 3.3, I distinguish between five affix

¹⁹For the remainder of this study, I will use “values” to refer to “morphosyntactic values”.

Table 19:
Affix position

	pfx	pfx + sfx	pfx + sfx + ifx	sfx	sfx + ifx
adjective	36	259	48	1365	0
noun	8	84	62	1436	2
verb	407	889	164	3093	8

positions found in the dataset. Table 19 shows the number of cells per part-of-speech expressed by markers in the five positions. For the analysis, I merged the two positions that include infixes, because the sfx + ifx category on its own has too few observations to allow for any meaningful insights. This leaves the following four values of affix position that are considered in the analysis: pfx, pfx+sfx, sfx, and has_ifx.

4.2

Modelling the probability of zero marking

To estimate the probability of zero marking in inflection, I aggregated the data by type of cell, language, and part-of-speech. This means that each datapoint corresponds to a proportion of zero marking (0.81) for a given type of cell (NOM.SG) in a given language (Hungarian) for a given part-of-speech (noun). As shown in Table 15, the dataset contains cells with proportions of zero marking that equal 0 or 1. Therefore I fitted a Bayesian zero-one-inflated beta regression model. Zero-one-inflated beta regression models consist of two components. The first component is the regular beta regression model, which deals with proportion values within the interval (0,1). The second component is a logistic regression component that estimates the probability of either of the extremes 0 or 1 as opposed to the proportion data within (0,1).

The models were fitted using Stan (Carpenter *et al.* 2017) with the *brms* package (Bürkner 2017) in R (R Core Team 2021). I additionally controlled for the phylogenetic relations between languages using a phylogenetic regression term, following the method described in Guzmán Naranjo and Becker 2022. This term does not model the relations between languages in a categorical way but includes the information of the entire phylogenetic tree and forces the estimates of individual languages to co-vary according to the tree.²⁰ In other words,

²⁰ The phylogenetic tree is taken from Glottolog (Hammarström *et al.* 2021). For details, see code-phylogeny.R in the supplementary materials.

if two languages share many nodes on the tree, the model forces their coefficients to be very similar. If two languages are not related at all, the model allows their estimates to vary freely. For instance, if five closely related languages have very high observed proportions of zero markers in a given cell, the model does not take those five observations as independent data points, but assigns much less confidence and/or lowers the predicted probability of zero marking in that cell.

The final model predicts the probability of zero marking from the part-of-speech, affix position, number of values per cell, number of lemmas, and orthographic representation. In addition, I used type of cell and phylogenetic relations between languages as group-level effects.²¹ Figures 3 and 4 show the conditional effects for the different predictors for the beta and the zero-one-inflation components, respectively.²² The points and solid lines correspond to the mean values of the posterior distributions; the error bars and error bands show the 95% credible interval. This approach allows a straightforward interpretation: given the data and the model, we can be 95% certain that the estimated values lie within that interval. Note that the three numerical predictors are all standardized, so that they have a mean of zero and a standard deviation of 1.

From Figure 3, we see that none of the predictors has a clear impact on the probability of zero marking within the interval (0,1). Across all predictors, the mean predictions lie between 0.15 and 0.3. The results thus show that the probability of zero marking to occur, excluding systematic absence or presence thereof, does not depend much on the predictors explored here. This does not necessarily mean that a better model is needed. It suggests that there is a high degree of idiosyncratic variation across languages, and that no clear association

²¹ To select a reasonable combination of predictors, I fitted several models and compared their performance using approximated leave-one-out cross-validation as described by Vehtari *et al.* (2017). Due to the low number of proportions of 1, I modelled conditional-one-inflation with an intercept-only model. See `code-prob.R` in the supplementary files for details on the conditional-one-inflation.

²² I only report the results of the model based on markers_A which allow for stem alternations. All conditional effects of the model based on markers without stem alternations can be found in `ce-probcheck-mu-<predictor>.pdf` and `ce-probcheck-zoi-<predictor>.pdf` in the supplementary materials.

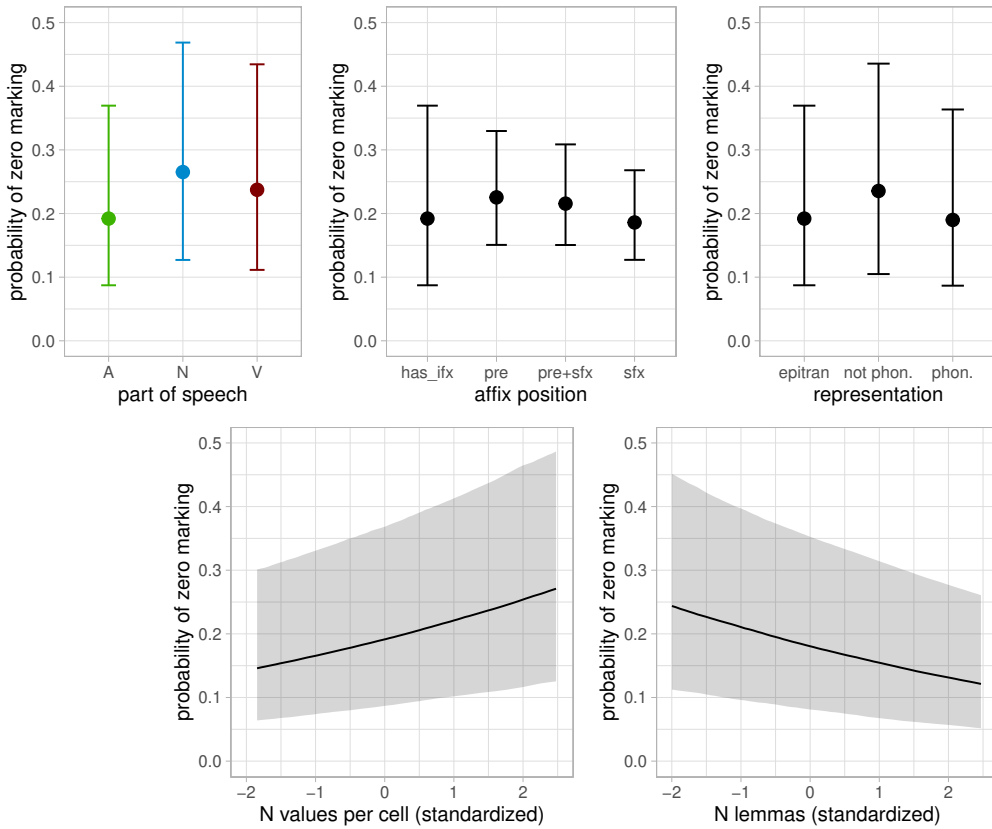


Figure 3: Conditional effects for the beta regression component

can be established with other relevant grammatical properties of the inflectional systems, at least not the ones tested here.

Figure 4 shows the model results for the zero-one-inflation component. It predicts the probability of a cell being exclusively zero marked or never zero marked, as opposed to probability values between those two extremes. As was shown in Table 15, no zero marking per cell is common in the data (6,437 markers out of 7,861), while exclusively zero marked cells are very rare (43 markers out of 7,861). This means that zero-one-inflation predictions largely correspond to the probability of no zero marking for a given cell. We can thus interpret the conditional effects shown in Figure 4 as the probability of the absence of zero marking. For the predictors part-of-speech, affix po-

Zero marking in inflection

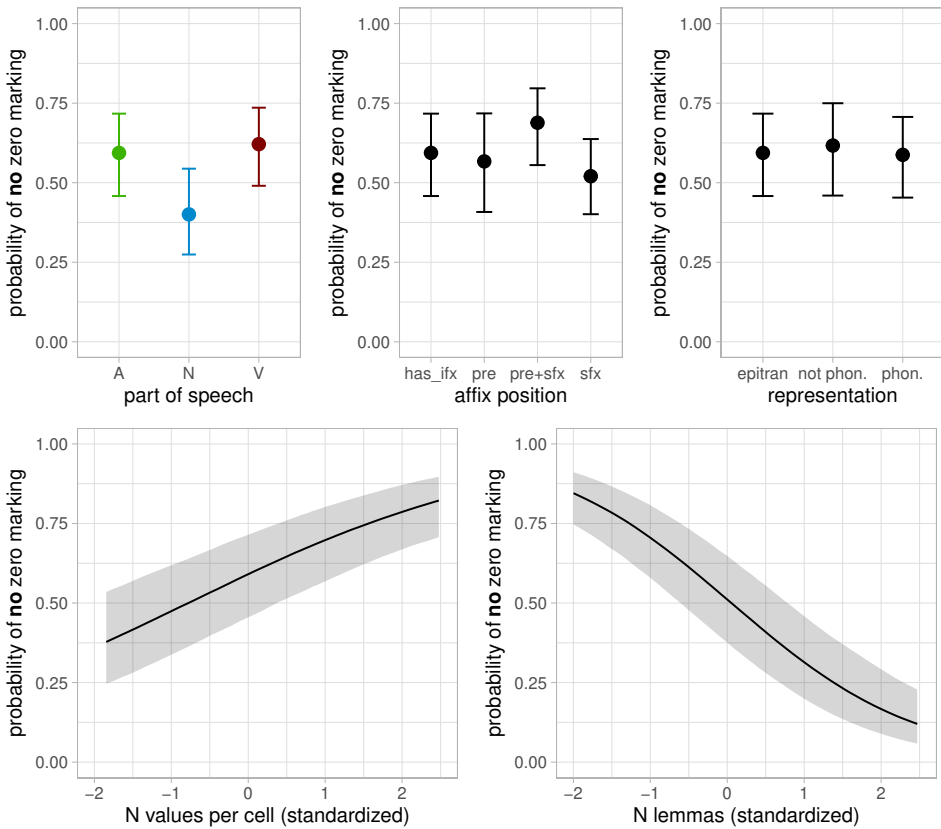


Figure 4: Conditional effects for the zero-one-inflation component

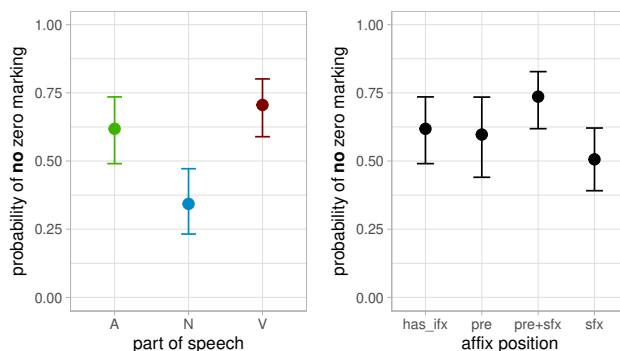
sition, and phonological representation, we find no substantial trends regarding a preference against zero marking. For part-of-speech, adjectives and verbs appear to have a slightly higher probability than nouns of avoiding zero marking altogether, but we have little certainty about this difference. The same can be said about the affix order pfx + sfx; it has a slightly higher tendency to avoid all zero marking than the other positions, but no clear picture emerges.

In contrast to the predictions from the beta component, we do find clear effects of the number of values per cell and the number of lemmas. The more lemmas are available, the lower the probability of encountering not a single case of zero marking. This is expected and shows that the number of lemmas needs to be controlled for. The

number of values per cell has a positive effect on the probability of avoiding zero marking altogether. While cells expressing fewer values show no strong preference for or against zero marking, the model predicts a strong preference against all zero marking for cells with many values. This does not restrict where zero marking is likely to occur, but it predicts the total absence of zero marking for complex cells, with a high probability of 0.8.

As mentioned in Section 3.6, it is important to consider the distribution of zero marking in morphomic paradigms as well. I fitted another Bayesian zero-one-inflated beta regression model using morphomic paradigms with the same predictors as described above. Only the predictors including information on cells (cell, number of values per cell) were no longer included. The predictions from the beta regression component are similar to those of the full paradigms, which is why I do not discuss them here in detail.²³ The overall predicted probability of zero marking is just below 0.2, which is slightly lower than in full paradigms. This suggests that zero marking is syncretic in a portion of the dataset. As the credible intervals are very wide in both models and overlap, we cannot be very certain about this finding. For the zero-one-inflation component of the model, the conditional effects of part-of-speech and affix position allow for additional insights. The model predictions for these two variables are shown in Figure 5. We

Figure 5:
Conditional
effects for the
zero-one-
inflation
component
of morphomic
paradigms



²³See the file `code-morphomic.R` for details. The conditional effects for all predictors of the model using morphomic paradigms are found in the supplementary materials as `ce-probmorph-mu-<predictor>.pdf` and `ce-probmorph-zoi-<predictor>.pdf`.

see that the patterns are similar, only the differences between parts-of-speech are much stronger now. With morphomic paradigms, we can be certain that verbs and adjectives have a stronger tendency than nouns to avoid zero marking altogether. The same holds for the affix position. Figure 5 shows that systems with prefixes and suffixes are more likely to avoid zero marking altogether than systems with suffixes only.

FUNCTIONS ASSOCIATED WITH ZERO MARKING 5

Cells with the highest probability of zero marking 5.1

To explore which cells are most likely to be zero marked, I subsetted the dataset to include only those cells with a proportion of zero forms ≥ 0.1 in at least 10% of the languages. Subsetting the data in such a way was necessary because of the high number of cell types. The threshold is a heuristic chosen to restrict the following analysis only to cells with a reasonable cross-linguistic probability of being expressed by zero markers. This step retains the 18 types of cells that show the strongest association with zero marking in the observed distributions.²⁴

In order to estimate the probability of zero marking in these cells, I fitted a Bayesian beta regression model that predicts the probability of zero marking from the type of cell.²⁵ In addition, I added the number of values per cell and lemmas as group-level intercepts as well as phylogenetic controls to account for phylogenetic biases in the data.

²⁴ The exact figures, including the number of languages per cell, are found in `cells-merged.csv` in the supplementary materials.

²⁵ In this case, I used beta regression instead of zero-one-inflated beta regression for a combined prediction from both processes. To do so, I converted proportions of zero to 0.0000001 and proportions of 1 to 0.9999999. Again, I compared several models using approximated leave-one-out-cross-validation. See `code-cells.R` in the supplementary materials for details.

Figure 6 shows the observed proportions of zero forms (black triangles) together with the model predictions (dots, error bars, and error bands).²⁶ Again, the dots represent the mean values of the posterior distribution of the zero probabilities, and the error bars and bands show the 95% credible intervals. The observed proportions of zero forms still differ across cells and parts-of-speech, ranging from 0.1 (2SG.PRS verb forms and DAT.SG adjectives) to above 0.7 (INDF.SG nouns). Although adjectives have fewer cells that met the threshold criteria than nouns and verbs, Figure 6 shows that the cells that do meet them have comparatively high proportions of zero marking. In nominal cells, we find a wider range, including the highest overall proportions of zero marking. Verbs show the lowest proportions of zero marking compared to the other parts-of-speech.

When comparing the results of the model with the observed proportions, the predicted probabilities of zero markers reflect the observed proportions, for the most part. The top plot in Figure 6 shows a few differences, though. For some cells, the predicted probability is much lower than their observed proportions, namely for PL.VOC in adjectives, as well as ACC.SG and INDF.SG in nouns. This points to a bias in the observed distributions, which is also reflected in the large credible intervals of the predictions. The PL.VOC cell is featured in four languages of the dataset, namely in Czech, Georgian, Irish, and Sanskrit. In this case, the high proportion of zero marking is mainly an artefact of the data. The PL.VOC cell is exclusively zero marked in the Czech data. Irish has a low proportion of zero marked PL.VOC cells (0.22), and Georgian as well as Sanskrit do not feature zero marking for the PL.VOC cells of adjectives. Thus, in this case, the high overall proportion largely comes from a single language, which is then adjusted to a much lower prediction in the model, together with large credible intervals to indicate the high level of uncertainty. A similar explanation applies to the ACC.SG cell in nouns. It is featured in 26 languages in the dataset, including phylogenetically unrelated languages. However, the higher observed proportion of zero marking is due to high proportions in a few, mostly related, languages with large

²⁶ All conditional effects of the model based on markers without stem alternations can be found in `ce-cells-check-<predictor>.pdf` in the supplementary materials.

Zero marking in inflection

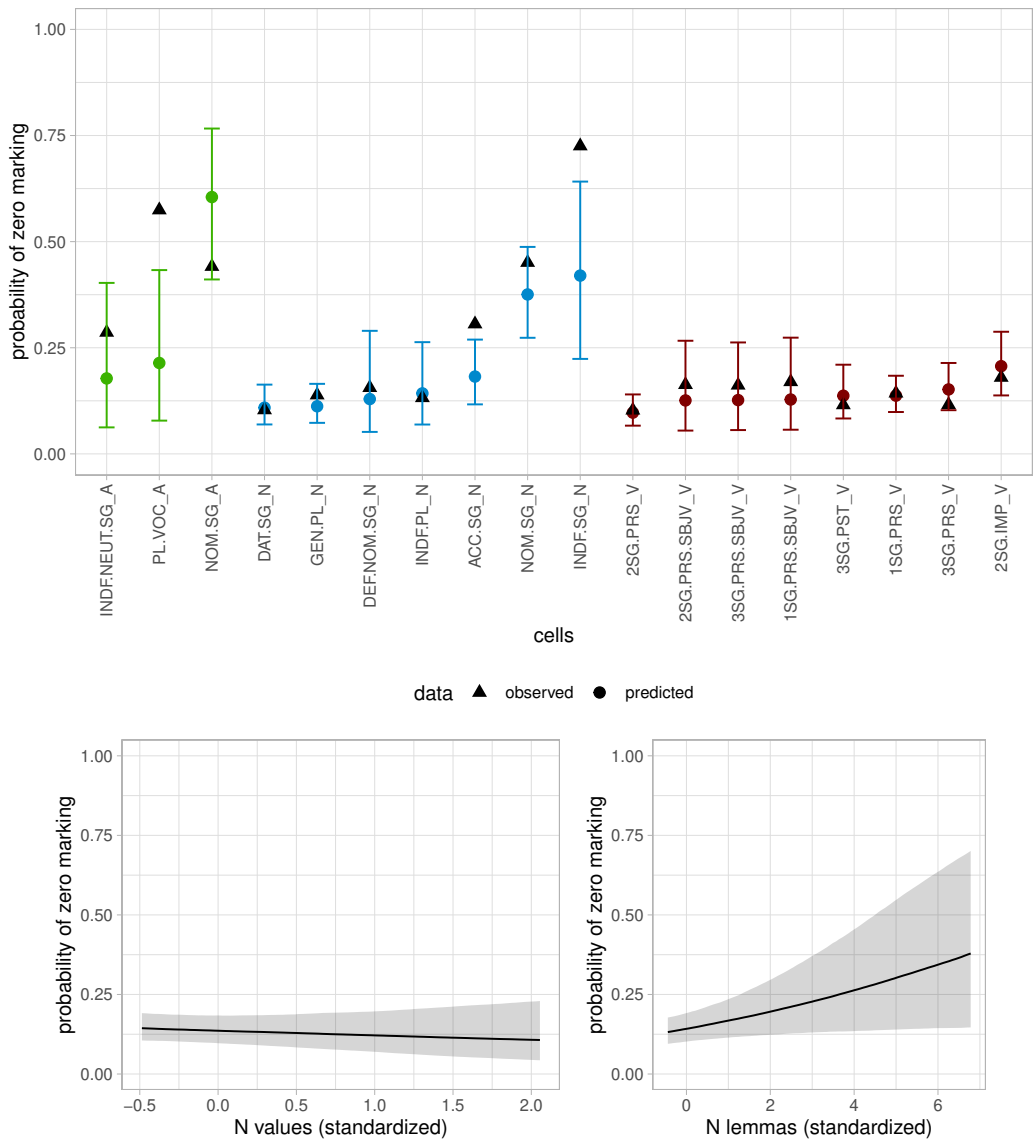


Figure 6: Conditional effects for cells most strongly associated with zero marking

datasets.²⁷ For the INDF.SG cell, the lower predicted probability of zero marking is also the consequence of a bias in the observed proportions. Here, the bias comes from Norwegian Bokmål, which makes up more than 50% of all observations for this cell, and which has a very high proportion (0.88) of zero marking.

Comparing the predictions across cells and parts-of-speech, we see that adjectival cells have a very high probability of being zero marked. This is noteworthy, as adjectives had only very few cells that met the threshold to begin with. While generally not associated with zero marking, the adjectival cells that are zero marked appear to be those with the strongest association with zero marking across parts-of-speech. Nominal cells are generally predicted to have lower probabilities of zero marking, except for the NOM.SG and the INDF.SG cells, which rank second and third for the predicted probability of zero marking. All verbal cells range between 0.1 and 0.25 for the probability of zero marking. The cell that stands out for having the highest probability of zero marking is the 2SG.IMP cell, which will be further discussed in Section 7.2.

5.2 *Values with the highest probability of zero marking*

The fact that the languages in the dataset differ with respect to the combinations of values in single cells makes it somewhat difficult to assess the association between zero marking and cells that are less common in the dataset. It is therefore important to consider the association of single grammatical values and zero marking as well. Note that, due to the way in which zero markers were extracted, pulling apart the values of cells and analyzing their association with zero marking does not translate directly into the traditional analysis of an abstract feature value, e.g. singular, as being zero marked. Rather, the singular value being expressed by a zero marker refers to all cells in the dataset that encode singular (potentially besides other feature values) and that are zero marked.

²⁷ This includes German (0.77), Old English (0.50), Finnish (0.37), Russian (0.35), Ukrainian (0.23), Polish (0.22), and Serbian-Croatian-Bosnian (0.30).

In order to examine the association of single values with zero marking, I applied a similar threshold heuristic as in Section 5.1 to select those values that show the strongest association with zero marking. I only included values with an overall proportion of zero marking ≥ 0.02 that are featured in 10% of the languages per part-of-speech. This led to the selection of 21 values in total.²⁸ To assess how robust the observed proportions of zero marking are, I fitted a Bayesian beta regression model, adding a phylogenetic control and the number of cells and lemmas as group-level effects.²⁹

Figure 7 shows the observed proportions (triangles) together with the model predictions (dots, lines).³⁰ The dots represent the mean values of the posterior distribution of the zero marker probabilities; error bars and bands indicate the 95% uncertainty intervals. The distributions in Figure 7 mostly mirror the tendencies seen in Figure 6 in the previous section. Almost all values that meet the threshold (and are thus the values with the highest proportions of zero marking) have also been part of the cells most likely to be zero marked. Only the nominal value VOC, and the verbal values PROG, PL, and NFIN have not been part of the cells most associated with zero marking. Compared to cells, values show much lower absolute proportions of zero marking. This is expected, since single values potentially occur in many different contexts, not all of which are necessarily zero marked. As for the three parts-of-speech, we now see the highest proportions for nominal values. Adjectival and verbal values show lower proportions of zero marking.

Turning to the model predictions, we see that in the case of values, the probability of zero marking is generally estimated by the model to be higher than the observed proportions. This can be explained by the fact that the model takes into account information on the affix position, the number of cells, and the number of lemmas. The effects of

²⁸The exact figures, including the number of languages per value, are found in `values-merged.csv` in the supplementary materials.

²⁹I used the same method as for the model described in Section 5.1. See `code-values.R` in the supplementary materials for details.

³⁰All conditional effects of the model based on markers without stem alternations can be found in `ce-values-check-<predictor>.pdf` in the supplementary materials.

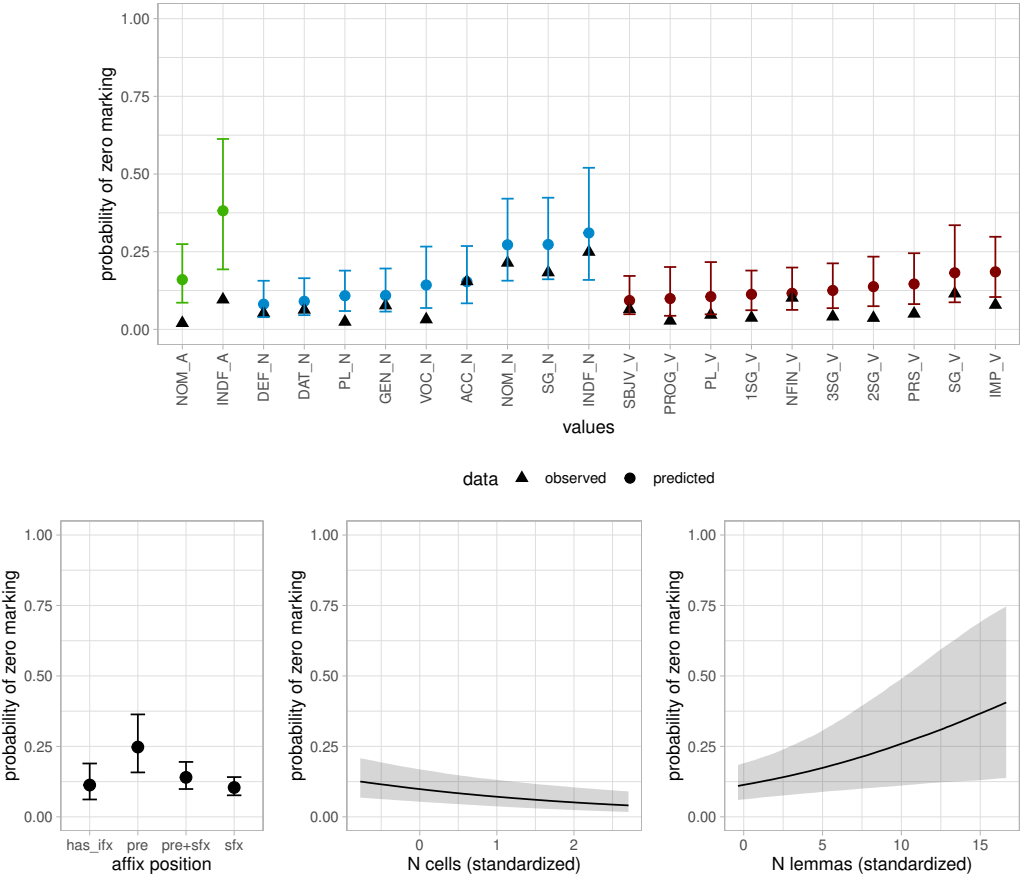


Figure 7: Conditional effects for the values most associated with zero marking.

single values thus correspond to their effects once all the other predictors are controlled for. Interestingly, the affix position is also relevant in this case. The model predicts a higher probability of zero marking for systems with prefixes as opposed to those with suffixes.

The highest predicted probabilities of zero marking are found for the indefinite value in adjectival and nominal inflection. This mirrors the model results of cells shown in Figure 6. Other values with a comparatively high probability of zero marking are SG and NOM for nouns, as well as IMP and SG for verbs. These results also reflect the tendencies seen with cells in Section 5.1.

THE FREQUENCY OF ZERO MARKERS IN LANGUAGE USE

6

To assess the usage frequencies of inflection markers and their phonological length including zero, I analyzed the distribution of zero markers in the Universal Dependencies treebanks (UD) (Zeman *et al.* 2023). To do this, I merged the adjective, noun, and verb forms in UniMorph identified as zero forms with the Universal Dependencies data. I only included the languages for which a phonological transcription was available, so that marker length could be approximated in a more realistic way. From the original dataset, 20 languages have phonological transcriptions and are represented in UD. When merging UniMorph forms with forms in UD, I did not include cell information, but merged the forms purely based on their orthographic representation. The identification of zero markers, however, was based on the phonological transcriptions and the `markerA` extraction, as described in Section 3. The resulting dataset contains 9,975 types of markers, which are made up of 51 types of zero markers (across different language and part-of-speech combinations) and 9,924 distinct types of overt markers. In terms of token frequencies, zero markers make up 23% of all the marker occurrences (7,382,497 tokens in total). For the purposes of this study, the distribution of zero and overt markers in UD

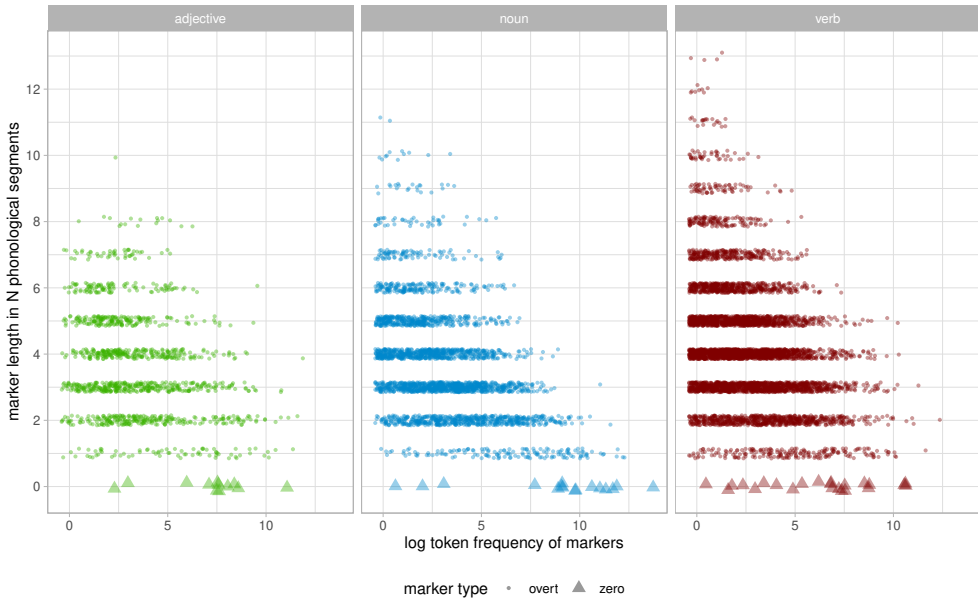


Figure 8: Association between marker token frequency and length

is measured by their log-transformed token frequencies.³¹ The length of the markers corresponds to the number of phonological segments identified with the UniMorph dataset. Figure 8 shows the relation between log token frequencies and marker length for adjectives, nouns and verbs. Overt markers are shown as dots, and zero markers are indicated by triangles. As expected, Figure 8 shows a consistent tendency across the three parts-of-speech for more frequent markers to

³¹ Frequency is but one of several possible measures of the distribution of linguistic expressions. Common alternatives are contextual probability and informativity (average contextual probability). Some studies suggest that these measures are more strongly associated with the length of an expression (e.g. Barth 2019; Cohen Priva 2015; Jurafsky *et al.* 2001; Levshina 2018; Piantadosi *et al.* 2011). However, which measure is “best” seems to depend on the corpus size and the phenomenon at hand. Given that there is no good suggestion from the literature as to which measure is most strongly associated with expression length in inflectional morphology, the present study uses frequency as a first, straightforward approach. Future research will be necessary to assess efficiency effects using other distribution measures.

be shorter. For less frequent markers, however, there does not seem to be a strong tendency to be longer; we also find many infrequent markers that are short. Figure 8 does not show any clear tendencies for zero markers either. For adjectives and nouns, they appear to have comparatively high frequencies, whereas no such trend is apparent for verbs.

To test the association shown in Figure 8, I fitted a Bayesian hurdle Poisson model, predicting the marker length from their frequencies. Similarly to the zero-one-inflated beta models, a hurdle Poisson model consists of two components. The Poisson component predicts count data, and the hurdle consists of a logistic regression component that predicts the probability of markers of length zero. This allows us to compare the effect of frequency on marker length between zero and overt markers.

In order to determine which predictors other than token frequency should be included, I fitted a series of 9 models that included different combinations of token frequency with part-of-speech, affix position, and number of cells. The performance of these models was then compared to select the final model. I used approximated leave-one-out cross-validation for the comparison, following the method described by Vehtari *et al.* (2017).³² The final model includes token frequency and affix position as well as their interaction and the phylogenetic control.

Figure 9 shows the conditional effects for the Poisson component, i.e. the part of the model that predicts the length of overt markers. We find a clear negative effect of marker frequency, confirming previous results from the literature. On average, low frequency markers are predicted to be about 0.15 phonological segments longer than high frequency markers. The position of the affix also proves relevant for marker length. Despite the effect being smaller, the model predicts a substantial difference in marker length between systems only using suffixes and all other systems. This becomes more evident when considering the interaction between token frequency and affix position. The effect of frequency is greater for systems using only suffixes than for all other systems, reaching an average difference of 0.25 phonological segments between low-frequency and high-frequency markers.

³² See code-ud.R in the supplementary materials for details.

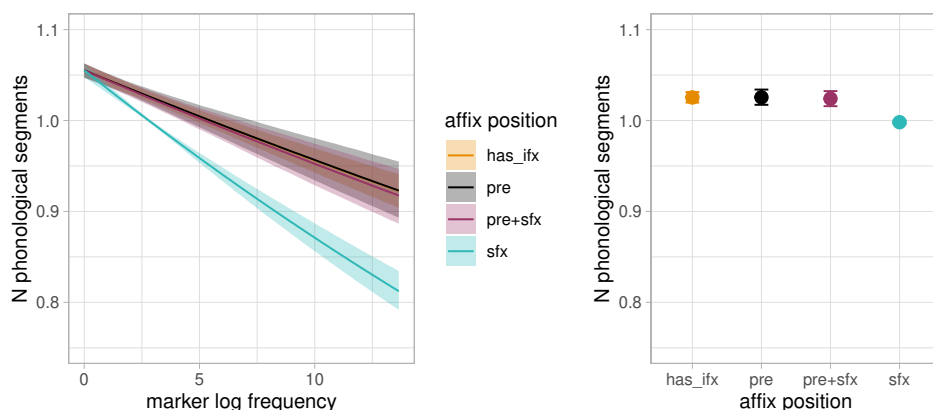


Figure 9: Conditional effects for the Poisson component

We can thus conclude that suffixes are more sensitive to the effect of marker frequency than the other types of affixes. We see the conditional effects for the hurdle component in Figure 10. They represent the effect that the predictors have on the probability of a zero marker occurring.

In stark contrast with the effects predicted for the phonological length of markers, neither token frequency nor affix position affect the probability of a zero marker. The small credible intervals show that this is not an issue of uncertainty or too few observations. We can be confident in the model results that, given the data, the probability of zero marking occurring is not associated with the token frequency of that marker or the affix position that the system uses. This means that there is indeed a clear difference between the effect of frequency on

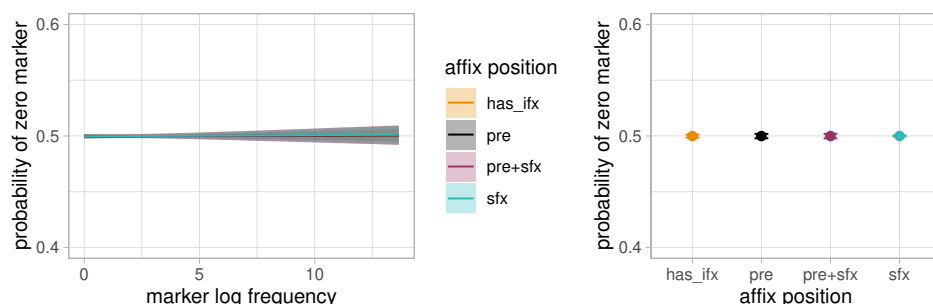


Figure 10: Conditional effects for the hurdle component

marker length in general and the occurrence of zero marking. Zero marking does not simply follow the general trend of marker length being associated with marker frequency.

DISCUSSION

7

The probability of zero marking

7.1

The results of this study allow for a number of important insights into cross-linguistic trends of zero marking in inflection. The model results predicting the probability of zero marking in inflectional paradigms (Section 4) showed three important points. First, zero marking generally affects adjectives, nouns, and verbs fairly equally, and the occurrence of zero marking is not sensitive to the affix position(s) used for inflection. The only notable difference across parts-of-speech and affix positions was found with the total absence of zero marking (zero-one-inflation component). Adjectives and verbs were more likely than nouns to avoid zero marking altogether. The same was seen for systems with prefixes and suffixes as opposed to systems with suffixes only. This effect was shown to be more pronounced when analyzing morphomic paradigms (cf. Figure 5), which are based on forms alone and where syncretic forms are counted only once per paradigm. As the overall probability of zero marking is rather low (0.1–0.3), zero marking is not a default strategy for inflection. This finding provides quantitative support for the proposal by Stolz and Levkovych (2019, 373), who argue that zero marking in inflection should be treated as a “morphological mismatch on a par with established categories such as suppletion and syncretism”. Zero marking is not a common strategy to encode inflection.

Second, we saw an effect of part-of-speech and affix position when analyzing zero marking in morphomic paradigms. Based on forms only, with no information about cells, zero marking was more likely to be absent altogether in adjectives and verbs as opposed to nouns. The same applied to systems with prefixes and suffixes as opposed to suffixes only. This does not mean that nouns and systems with suffixes

have a stronger preference for zero marking. It rather suggests that the complete absence of zero marking is less likely in those cases.

Third, an increasing number of values per cell was shown to be a strong predictor for a high probability of zero marking being avoided altogether. The predictor number of values per cell can be taken to quantify how semantically complex a marker is. The fact that more complex cells strictly avoid zero marking is reminiscent of what has been discussed as isomorphism or iconicity in the literature (cf. Haspelmath 2008b; Lehmann 1974; Downing and Stiebels 2012; Givón 1991). While approaches differ in their details, the general idea is that the complexity or amount of linguistic structure reflects the complexity or amount of functional structure (meaning). It remains an open question, however, whether the number of morphosyntactic values per cell reflects functional complexity in the first place, and what the functional motivation for any such effect might be. It is likely that usage distributions and frequencies are a confounding factor, in that cells expressing more values may also be cells that are used less frequently. Their preference for longer markers could thus be a consequence of frequency rather than some iconicity principle.

7.2

Cells and values associated with zero marking

Sections 5.1 and 5.2 focused on a selection of cells and morphosyntactic values and their association with zero marking. The results showed that even though zero marking exhibits a high degree of variation across lemmas and languages, it is not distributed randomly across inflectional paradigms. Some cells and values are comparatively likely to be zero marked across languages. For adjectives and nouns, INDF, NOM, and SG (and cell combinations thereof) were the values with the highest predicted probability of zero marking. For verbs, the probabilities of zero marking tended to be lower. The values of IMP, SG, 3, and PRS (and cell combinations thereof) stood out as those with the highest probability of zero marking. The NOM.SG cell for adjectives was the only cell for which the probability of zero marking was predicted to be above 0.5. In other words, this is the only cell for which we can expect zero marking to be more likely than overt marking. In all other cases, predicted probabilities

were well below 0.5. This means that the vast majority of inflectional marking is in fact overt, and zero marking is more of an exception.

The values of NOM and SG, as well as their combination, have long been associated with zero marking in the typological literature (e.g. Croft 2003; Greenberg 1963, 1966; Haspelmath and Karjus 2017; Haspelmath 2021; Jakobson 1983[1939]; Koch 1995). Interestingly, there is less discussion in the literature about zero marking of the INDF value, which showed the strongest trend towards zero marking in this study. Two verbal values that have been related to zero marking in the literature are third person (Bickel *et al.* 2015; Cysouw 2003; Siewierska 2010) and present tense (Bybee and Dahl 1989, 55; Bybee 1994, 248). The results of this study confirm the association. Although neither values show a cross-linguistic preference towards being zero marked, they are part of the values with the highest probabilities of zero marking.

Imperatives, especially 2SG forms, have also been mentioned in the literature as being prone to zero marking (e.g. Aikhenvald 2010; Croft 2003; Greenberg 1966; Haspelmath 2021; Koch 1995; Siewierska 2010). The results of the present study thus fit well with the expectations from the literature. Instead of phonetic reduction, previous studies have argued for a functionally motivated non-development scenario for zero marking in (2SG) imperatives. The idea is that the second person is highly recoverable in imperative contexts, e.g. as opposed to contexts of indicative verb forms. Thus, on the level of syntax, many languages allow or require the use of imperatives with no overt second person subject pronoun. This in turn means that the source construction of a verbal person marker is often not available for imperative forms (Aikhenvald 2010, 147; Nikolaeva 2007, 163; Sadock and Zwicky 1985, 173). The cross-linguistically common absence of a suitable source construction for person markers in imperative contexts may thus ultimately account for the high probability of zero marking, especially for person-number agreement values. In addition, the use of bare verb forms for imperatives has been motivated by iconicity (Aikhenvald 2010, 46). According to her, using the shortest verb form makes imperatives very direct and abrupt. This can convey urgency and reflect that imperatives usually call for an immediate reaction.

7.3

Frequency effects and affix position

Section 6 examined the association between the token frequency of inflection markers and their length, including zero marking. For overt inflectional markers, the present study provided further evidence of Zipfian effects. Markers with a higher log frequency were predicted to have shorter forms (i.e. number of phonological segments). This corroborates previous findings about form-frequency effects for inflectional markers (cf. Haspelmath and Karjus 2017; Stave *et al.* 2021).

An aspect that has not so far been addressed in quantitative corpus studies is the effect that the position of the inflection marker has. The results from this study showed a clear difference between inflectional systems using only suffixes and those that use different combinations of prefixes, suffixes, and infixes. If inflectional markers are strictly suffixes, their length is predicted to be shorter than if the system uses a combination of affix positions. The effect of token frequency on marker length was also shown to be stronger for suffixes than for other combinations of marker positions. This means that suffixes are more susceptible to frequency effects on marker length than other affix positions are.

A potential explanation for this difference across affix positions is phonetic reduction over time. We know from the literature that phonetic material at the end of words is reduced at higher rates than material at the beginning of words (Bybee *et al.* 1990, 19; Hall 1988). There is also evidence for word-initial (or domain-initial) syllables to be more prominent than other syllables (e.g. Beckman 1998; Smith 2005; Cho *et al.* 2007; Kim 2004; Keating *et al.* 2004). Especially word-initial consonants tend to be strengthened and lengthened (e.g. White *et al.* 2020; Cho and Keating 2009; Fougeron 2001; Cho and Keating 2001). This is relevant, since Bybee *et al.* (1990, 26) find that inflectional prefixes are cross-linguistically significantly more likely to have initial consonants than inflectional suffixes. Taken together, it is plausible that these properties contribute to suffixes being more likely candidates for phonetic reduction over time than affixes in other positions.

*Support for the non-development scenario
of zero markers*

7.4

The other major finding from Section 6 is that the association between token frequency and marker length did not hold for zero markers. Their distributions in the Universal Dependencies treebanks showed that neither token frequency nor affix position were associated with the occurrence of a zero marker. This is evidence against the traditional (implicit) assumption in typology that zero markers behave like short markers in terms of their distribution in language use (e.g. Bybee 2011; Croft 2003; Greenberg 1966; Haspelmath 2021). At the same time, the results from this study confirm previous studies, arguing that coding efficiency and frequency may not be suitable or a sufficient explanation for zero marking in inflectional morphology (Stolz and Levkovych 2019; Guzmán Naranjo and Becker 2021; Bickel *et al.* 2015; Cysouw 2003; Siewierska 2010; Seržant and Moroz 2022).

The difference between overt and zero markers in terms of their association with token frequencies also provides evidence for the non-development scenario leading to zero markers. The other potential mechanism leading to zero marking is phonetic reduction. Phonetic reduction is commonly invoked as the mechanism responsible for the shortening of forms and the development of zero forms (Bybee 2003, 2007, 2015; Givón 2018; Haspelmath 2008a; Lehmann 2015). Bybee (2003, 2015) in particular has argued for phonetic reduction being a consequence of the repetition and automatization in production in the course of grammaticalization.

The main alternative to phonetic reduction is the differential non-development of a marker (cf. Bybee 1994; Cristofaro 2019, 2021; Haspelmath 2008a). For instance, we can imagine a scenario in which number is not marked on nouns at a given point in time. For independent reasons, plural marking could be developed. At the same time that the plural marker develops into an inflectional exponent, its absence becomes more systematically associated with the singular. Then, at some point, the singular is expressed by a zero form. In such a scenario, the zero marker results from the opposition to another new exponent in a different cell of the paradigm.

We can assume that phonetic reduction is at least in part responsible for the patterns found with overt markers, since we found a strong

association between token frequency and marker length. Given that such an effect was not found for zero markers, the role of phonetic reduction as the main factor driving their development is questionable. As was mentioned above, the other main mechanism that can lead to the development of zero marking is the differential non-development of an inflection marker. For such a scenario, usage token frequencies may still play a role, but much more indirectly. In a non-development scenario, the zero marker is merely a consequence of the development of a different marker. The development process thus depends on a number of factors that are not directly related to the zero marker itself. The results from Section 6 cannot offer direct evidence in favor of the non-development scenario, but they are more compatible with this scenario than with the phonetic reduction scenario. There is certainly no single answer as to which mechanism leads to zero marking; it is likely that both these mechanisms and others are involved, although probably to differing degrees. Diachronic corpus work is needed to shed more light on the development of zero marking and its cross-linguistic tendencies.

8

CONCLUSION

This study offers the first token-based overview of zero marking in adjectival, nominal, and verbal inflectional morphology across languages. Using the UniMorph dataset, it takes into account the behavior of single lemmas to capture variation across inflection classes and irregular forms. Regarding the probability of zero marking in inflection, the results showed that zero marking is generally not a preferred marking strategy, as it is predicted to occur in only 10–30% of inflected forms. No single cells or values showed a strong association with zero marking. Nevertheless, the values with the highest probability of zero marking (NOM, SG, INDF, 3, PRS, IMP) confirmed earlier observations from the typological literature. The findings further evidenced a high degree of idiosyncratic variation across languages and lemmas in the distribution of zero markers.

In addition, the study analyzed the token frequencies of zero markers together with those of overt markers in several corpora from

the Universal Dependencies treebanks. For overt markers, the results showed that the token frequency has a stronger effect on the phonological length of suffixes compared to other affixes. This fits into a broader picture of phonetic differences between suffixes and other positions. For the probability of zero markers, however, no association with their frequency was found. This is new evidence for a fundamental difference between the distribution of overt and zero markers. Zero markers do not simply follow the distributional patterns of short markers. This difference supports a differential non-development scenario of zero marking, rather than a phonetic reduction scenario.

ABBREVIATIONS

1 – first person, 2 – second person, 3 – third person, ABL – ablative, ACC – accusative, ALL – allative, AOR – aorist, COM – comitative, COND – conditional, DAT – dative, DEF – definite, EQTV – equative, ESS – essive, F – feminine, FRML – formal case, FUT – future, GEN – genitive, ILL – illative, IMP – imperative, IMPF – imperfect, INESS – inessive, IND – indicative, INDF – indefinite, INSTR – instrumental, IPFV – imperfective, M – masculine, N – neuter, NFIN – non-finite, NOM – nominative, ON – surface, PFV – perfective, PL – plural, PROG – progressive, PRP – purposive, PRS – present, PST – past, PTCP – participle, SG – singular, SBJV – subjunctive, TERM – terminative, VOC – vocative

REFERENCES

- Alexandra AIKHENVALD (2010), *Imperatives and commands*, Oxford University Press, Oxford.
- Sergiu AL-GEORGE (1967), The semiosis of zero according to Pāṇini, *East and West*, 17(1):115–124.
- Stephen R. ANDERSON (1992), *A-morphous morphology*, Cambridge University Press, Cambridge.

- Peter ARKADIEV (2016), Возможны ли однопадежные системы? [Are monocausal systems possible?], in Józefina PIĄTKOWSKA and Gennadij ZELDOWICZ, editors, *Znaki czy nie znaki? – II. zbiór prac lingwistycznych*, pp. 9–37, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa.
- Matthew BAERMAN, Dunstan BROWN, and Greville G. CORBETT (2017), *Morphological complexity*, Cambridge University Press, Cambridge.
- Matthew BAERMAN and Greville CORBETT (2012), Stem alternations and multiple exponence, *Word Structure*, 5(1):52–68, doi:10.3366/word.2012.0019.
- Marianne BAKRÓ-NAGY (2022), Consonant gradation, in Marianne BAKRÓ-NAGY, Johanna LAAKSO, and Elena SKRIBNIK, editors, *The Oxford guide to Uralic languages*, pp. 859–867, Oxford University Press, Oxford.
- Sebastian BANK and Jochen TROMMER (2015), Learning and the complexity of Ø-marking, in Matthew BAERMAN, Dunstan BROWN, and Greville CORBETT, editors, *Understanding and measuring morphological complexity*, Oxford University Press, Oxford.
- Danielle BARTH (2019), Effects of average and specific context probability on reduction of function words BE and HAVE, *Linguistics Vanguard*, 5(1):20180055, doi:10.1515/lingvan-2018-0055.
- Jill BECKMAN (1998), *Positional faithfulness*, Ph.D. thesis, University of Massachusetts, Amherst.
- Sacha BENIAMINE and Matías GUZMÁN NARANJO (2021), Multiple alignments of inflectional paradigms, *Proceedings of the Society for Computation in Linguistics*, 4:216–227, doi:10.7275/ymc0-p491.
- Balthasar BICKEL, Alena WITZLACK-MAKAREVICH, Taras ZAKHARKO, and Giorgio IEMMOLO (2015), Exploring diachronic universals of agreement: Alignment patterns and zero marking across person categories, in Jürg FLEISCHER, Elisabeth RIEKEN, and Paul WIDMER, editors, *Agreement from a diachronic perspective*, pp. 29–52, De Gruyter, Berlin.
- James BLEVINS (2003), Stems and paradigms, *Language*, 79(4):737–767.
- James BLEVINS (2005), Word-based declensions in Estonian, in Geert BOOIJ and Jaap VAN MARLE, editors, *Yearbook of Morphology 2005*, pp. 1–25, Springer, Dordrecht, doi:10.1007/1-4020-4066-0_1.
- James BLEVINS (2006), Word-based morphology, *Journal of Linguistics*, 42(3):531–573, doi:10.1017/S0022226706004191.
- James BLEVINS (2016), *Word and paradigm morphology*, Oxford University Press, Oxford.
- Bernard BLOCH (1947), English verb inflection, *Language*, 23(4):399–418, doi:10.2307/410300.
- Leonard BLOOMFIELD (1933), *Language*, Holt, New York.

- Olivier BONAMI (2012), Stems in inflection and lexeme formation, *Word Structure*, 5(1), doi:10.3366/word.2012.0016.
- Olivier BONAMI and Sacha BENIAMINE (2021), Leaving the stem by itself, in Sedigheh MORADI, Marcia HAAG, Janie REES-MILLER, and Andrija PETROVIC, editors, *All things morphology: Its independence and its interfaces*, pp. 81–98, Benjamins, Amsterdam, doi:10.1075/cilt.353.05bon.
- Sami BOUDELAA and William D MARSLÉN-WILSON (2001), Morphological units in the Arabic mental lexicon, *Cognition*, 81(1):65–92, doi:10.1016/S0010-0277(01)00119-6.
- Gilles BOYÉ and Gauvin SCHALCHI (2016), The status of paradigms, in Andrew HIPPISEY and Gregory STUMP, editors, *The Cambridge handbook of morphology*, pp. 206–234, Cambridge University Press, Cambridge.
- Dunstan BROWN (1998), Stem indexing and morphonological selection in the Russian verb: A network morphology account, in Ray FABRI, Albert ORTMANN, and Teresa PARODI, editors, *Models of inflection*, pp. 196–224, Niemeyer.
- Joan BYBEE (1994), The grammaticization of zero: Asymmetries in tense and aspect systems, in William PAGLIUCA, editor, *Perspectives on grammaticalization*, pp. 235–254, Benjamins, Amsterdam.
- Joan BYBEE (2003), Mechanisms of change in grammaticization: The role of frequency, in Brian JOSEPH and Richard JANDA, editors, *Handbook of historical linguistics*, pp. 602–623, Blackwell, Oxford.
- Joan BYBEE (2007), *Frequency of use and the organization of language*, Oxford University Press, Oxford.
- Joan BYBEE (2011), Markedness, in Jae Jung SONG, editor, *The Oxford handbook of typology*, pp. 1–11, Oxford University Press, Oxford.
- Joan BYBEE (2015), *Language change*, Cambridge University Press, Cambridge.
- Joan BYBEE and Östen DAHL (1989), The creation of tense and aspect systems in the languages of the world, *Studies in Language*, 13(1):51–103, doi:10.1075/sl.13.1.03byb.
- Joan BYBEE, William PAGLIUCA, and Revere PERKINS (1990), On the asymmetries in the affixation of grammatical material, in William CROFT, Suzanne KEMMER, and Keith DENNING, editors, *Studies in typology and diachrony. Papers presented to Joseph H. Greenberg on his 75th birthday*, pp. 1–42, Benjamins, Amsterdam.
- Paul-Christian BÜRKNER (2017), Brms: An R package for Bayesian multilevel models using Stan, *Journal of Statistical Software*, 80(1):1–28, doi:10.18637/jss.v080.i01.
- Bob CARPENTER, Andrew GELMAN, Matthew HOFFMAN, Daniel LEE, Ben GOODRICH, Michael BETANCOURT, Marcus BRUBAKER, Jiqiang GUO, Peter LI,

- and Allen RIDDELL (2017), Stan: A probabilistic programming language, *Journal of Statistical Software*, 76(1):1–32, doi:10.18637/jss.v076.i01.
- Taehong CHO and Patricia KEATING (2001), Articulatory and acoustic studies on domain-initial strengthening in Korean, *Journal of Phonetics*, 29(2):155–190, doi:10.1006/jpho.2001.0131.
- Taehong CHO and Patricia KEATING (2009), Effects of initial position versus prominence in English, *Journal of Phonetics*, 37(4):466–485, doi:10.1016/j.wocn.2009.08.001.
- Taehong CHO, James MCQUEEN, and Ethan COX (2007), Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English, *Journal of Phonetics*, 35(2):210–243, doi:10.1016/j.wocn.2006.03.003.
- Uriel COHEN PRIVA (2015), Informativity affects consonant duration and deletion rates, *Laboratory Phonology*, 6(2):243–278, doi:10.1515/lp-2015-0008.
- Matt COLER (2015), Aymara inflection, in Matthew BAERMAN, editor, *The Oxford handbook of inflection*, pp. 1–30, Oxford University Press, Oxford.
- Matt COLER (2018), Subtractive morphology & disfixation in Aymara case, in *Case and Agreement in Panará (... and Beyond)*, pp. 1–9, University of Groningen, Groningen, doi:10.13140/RG.2.2.26153.03682.
- Ellen CONTINI-MORAVA (2006), The difference between zero and nothing: Swahili noun class prefixes 5 and 9/10, in Joseph DAVIS, Radmila GORUP, and Nancy STERN, editors, *Advances in Functional Linguistics*, pp. 211–222, Benjamins, Amsterdam.
- Greville CORBETT (2007), Canonical typology, suppletion, and possible words, *Language*, 83(1):8–42, doi:10.1353/lan.2007.0006.
- Sonia CRISTOFARO (2019), Taking diachronic evidence seriously: Result-oriented vs. source-oriented explanations of typological universals, in Karsten SCHMIDTKE-BODE, Natalia LEVSHINA, Susanne Maria MICHAELIS, and Ilya SERŽANT, editors, *Explanation in typology: Diachronic sources, functional motivations and the nature of the evidence*, pp. 25–46, Language Science Press, Berlin.
- Sonia CRISTOFARO (2021), Typological explanations in synchrony and diachrony: On the origins of third person zeroes in bound person paradigms, *Folia Linguistica*, 55(s42-s1):25–48, doi:10.1515/flin-2021-2013.
- William CROFT (2003), *Typology and universals*, Cambridge University Press, Cambridge, 2nd edition.
- Michael CYSOUW (2003), *The paradigmatic structure of person marking*, Oxford University Press, Oxford.
- Eystein DAHL and Antonio FÁBREGAS (2018), Zero morphemes, in Rochelle LIEBER, editor, *Oxford research encyclopedia of linguistics*, pp. 1–30, Oxford University Press, Oxford, doi:10.1093/acrefore/9780199384655.013.592.

Ferdinand DE SAUSSURE (1916), *Cours de linguistique générale*, in Bally, Charles and Albert Sechehaye, editors, Payot, Lausanne.

Catharine DIEHL (2008), The empty space in structure: Theories of the zero from Gauthiot to Deleuze, *Diacritics*, 38(3):93–119, <https://www.jstor.org/stable/20616535>.

Holger DIESSEL (2019), *The grammar network: How linguistic structure is shaped by language use*, Cambridge University Press, Cambridge.

Laura J. DOWNING and Barbara STIEBELS (2012), Iconicity, in Jochen TROMMER, editor, *The morphology and phonology of exponence*, Oxford Studies in Theoretical Linguistics, Oxford University Press, Oxford.

Cécile FOUGERON (2001), Articulatory properties of initial segments in several prosodic constituents in French, *Journal of Phonetics*, 29(2):109–135, doi:10.1006/jpho.2000.0114.

Talmy GIVÓN (1991), Isomorphism in the grammatical code: Cognitive and biological considerations, *Studies in Language*, 15(1):85–114, doi:10.1075/sl.15.1.04giv.

Talmy GIVÓN (2018), *On understanding grammar*, Benjamins, Amsterdam.

Joseph GREENBERG (1963), Some universals of grammar with particular reference to the order of meaningful elements, in Joseph GREENBERG, editor, *Universals of language*, pp. 73–113, MIT Press, Cambridge, MA.

Joseph GREENBERG (1966), *Language universals: With special reference to feature hierarchies*, Mouton, The Hague.

Matías GUZMÁN NARANJO and Laura BECKER (2021), Coding efficiency in nominal inflection: Expectedness and type frequency effects, *Linguistics Vanguard*, 7(s3):20190075, doi:10.1515/lingvan-2019-0075.

Matías GUZMÁN NARANJO and Laura BECKER (2022), Statistical bias control in typology, *Linguistic Typology*, 26(3):605–670, doi:10.1515/lingty-2021-0002.

William HAAS (1957), Zero in linguistics description, in John Rupert FIRTH, editor, *Studies in linguistic analysis*, pp. 33–53, Blackwell, Oxford.

Christopher HALL (1988), Integrating diachronic and processing principles in explaining the suffixing preference, in John HAWKINS, editor, *Explaining language universals*, pp. 321–349, Basil Blackwell, London.

Harald HAMMARSTRÖM, Robert FORKEL, Martin HASPELMATH, and Sebastian BANK (2021), *Glottolog 4.4*, Max Planck Institute for the Science of Human History, Leipzig, <http://glottolog.org>.

Martin HASPELMATH (2008a), Creating economical morphosyntactic patterns in language change, in Jeff GOOD, editor, *Linguistic universals and language change*, pp. 185–214, Oxford University Press, Oxford.

- Martin HASPELMATH (2008b), Frequency vs. iconicity in explaining grammatical asymmetries, *Cognitive Linguistics*, 19(1):1–33, doi:10.1515/COG.2008.001.
- Martin HASPELMATH (2008c), A frequentist explanation of some universals of reflexive marking, *Linguistic Discovery*, 6(1):40–63, doi:10.1349/PS1.1537-0852.A.331.
- Martin HASPELMATH (2018), How comparative concepts and descriptive linguistic categories are different, in Daniël VAN OLMEN, Tanja MORTELMANS, and Frank BRISARD, editors, *Aspects of linguistic variation*, pp. 83–114, De Gruyter, Berlin, doi:10.1515/9783110607963-004.
- Martin HASPELMATH (2021), Explaining grammatical coding asymmetries: Form–frequency correspondences and predictability, *Journal of Linguistics*, pp. 1–29, doi:10.1017/S0022226720000535.
- Martin HASPELMATH, Andreea CALUDE, Michael SPAGNOL, Heiko NARROG, and Elif BAMYACI (2014), Coding causal–noncausal verb alternations: A form–frequency correspondence explanation, *Journal of Linguistics*, 50(3):587–625, doi:10.1017/S0022226714000255.
- Martin HASPELMATH and Andres KARJUS (2017), Explaining asymmetries in number marking: Singulatives, pluratives, and usage frequency, *Linguistics*, 55(6):1213–1235, doi:10.1515/ling-2017-0026.
- George HEWITT (1995), *Georgian: A structural reference grammar*, Benjamins, Amsterdam.
- Jane HILL and Ofelia ZEPEDA (1998), Tohono O’odham (Papago) Plurals, *Anthropological Linguistics*, 40(1):1–42.
- Charles HOCKETT (1967), The Yawelmani basic verb, *Language*, 43(1):208–222, doi:10.2307/411395.
- Roman JAKOBSON (1983[1939]), Zero sign, in Linda WAUGH and Morris HALLE, editors, *Russian and Slavic grammar: Studies 1931-1981*, pp. 1–14, De Gruyter, New York.
- Daniel JURAFSKY, Alan BELL, Michelle GREGORY, and William RAMOND (2001), Probabilistic relations between words: Evidence from reduction in lexical production, in Joan BYBEE and Paul HOPPER, editors, *Frequency and the emergence of linguistic structure*, pp. 229–254, Benjamins, Amsterdam.
- Patricia KEATING, Taehong CHO, Fougeron CECILE, and Chai-Shune HSU (2004), Domain-initial strengthening in four languages, in John LOCAL, Richard ODGEN, and Rosalind TEMPLE, editors, *Phonetic interpretation. Papers in laboratory phonology VI*, pp. 145–163, Cambridge University Press, Cambridge.
- Sahyang KIM (2004), *The role of prosodic phrasing in Korean word segmentation*, Ph.D. thesis, University of California, Los Angeles, https://linguistics.ucla.edu/wp-content/uploads/2021/11/SahyangKim_dissertation.pdf.

Harold KOCH (1995), The creation of morphological zeros, in Geert BOOIJ and Jaap VAN MARLE, editors, *Yearbook of Morphology 1994*, pp. 31–731, Springer, Dordrecht.

Christian LEHMANN (1974), Isomorphismus im sprachlichen Zeichen [Isomorphism in the linguistic sign], in *Linguistic Workshop II: Arbeiten Des Kölner Universalienprojekts 1973/4*, pp. 98–123, Fink, München.

Christian LEHMANN (2015), *Thoughts on grammaticalization*, Language Science Press, Berlin.

Natalia LEVSHINA (2018), Probabilistic grammar and constructional predictability: Bayesian generalized additive models of help + (to) Infinitive in varieties of web-based English, *Glossa*, 3(1):1–22, doi:10.5334/gjgl.294.

Natalia LEVSHINA (2022), *Communicative efficiency: Language structure and use*, Cambridge University Press, Cambridge.

Martin MAIDEN (1992), Irregularity as a determinant of morphological change, *Journal of Linguistics*, 28(2):285–312, doi:10.1017/S0022226700015231.

Peter Hugoe MATTHEWS (1972), *Inflectional morphology: A theoretical study based on aspects of Latin verb conjugation*, Cambridge University Press.

Arya D. MCCARTHY, Christo KIROV, Matteo GRELLA, Amrit NIDHI, Patrick XIA, Kyle GORMAN, Ekaterina VYLOMOVA, Sabrina J. MIELKE, Garrett NICOLAI, Miikka SILFVERBERG, Timofey ARKHANGELSKIY, Nataly KRIZHANOVSKY, Andrew KRIZHANOVSKY, Elena KLYACHKO, Alexey SOROKIN, John MANSFIELD, Valts ERNSTREITS, Yuval PINTER, Cassandra L. JACOBS, Ryan COTTERELL, Mans HULDEN, and David YAROWSKY (2020), UniMorph 3.0: Universal Morphology, in Nicoletta CALZOLARI, Frédéric BÉCHET, Philippe BLACHE, Khalid CHOUKRI, Christopher CIERI, Thierry DECLERCK, Sara GOGGI, Hitoshi ISAHARA, Bente MAEGAARD, Joseph MARIANI, Hélène MAZO, Asuncion MORENO, Jan ODIJK, and Stelios PIPERIDIS, editors, *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 3922–3931, European Language Resources Association, Marseille, France, <https://aclanthology.org/2020.lrec-1.483>.

William MCGREGOR (2003), The nothing that is, the zero that isn't, *Studia Linguistica*, 57(2):75–119, doi:10.1111/1467-9582.00100.

Georg Friedrich MEIER (1961), *Das Zéro-Problem in der Linguistik. Kritische Untersuchungen zur strukturalistischen Analyse der Relevanz sprachlicher Form*, Akademie Verlag, Berlin.

Igor MEL'ČUK (1994), Suppletion: Toward a logical analysis of the concept, *Studies in Language*, 18(2):339–410, doi:10.1075/sl.18.2.03mel.

Igor MEL'ČUK (2002), Towards a formal concept zero linguistic sign: Applications in typology, in Sabrina BENDJABALLAH, Wolfgang DRESSLER, Oskar PFEIFFER, and Maria VOEIKOVA, editors, *Morphology 2000: Selected*

papers from the 9th Morphology Meeting, Vienna, 24–28 February 2000, pp. 241–258, Benjamins, Amsterdam.

Marianne MITHUN (1986), When zero isn't there, *Annual Meeting of the Berkeley Linguistics Society*, 12(0):195–211, doi:10.3765/bls.v12i0.1882.

Fabio MONTERMINI and Olivier BONAMI (2013), Stem spaces and predictability in verbal inflection, *Lingue e linguaggio*, 2:171–190, doi:10.1418/75040.

David R. MORTENSEN, Siddharth DALMIA, and Patrick LITTELL (2018), Epitran: Precision G2P for many languages, in Nicoletta CALZOLARI, Khalid CHOUKRI, Christopher CIERI, Thierry DECLERCK, Sara GOGGI, Koiti HASIDA, Hitoshi ISAHARA, Bente MAEGAARD, Joseph MARIANI, Hélène MAZO, Asuncion MORENO, Jan ODLJK, Stelios PIPERIDIS, and Takenobu TOKUNAGA, editors, *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 2710–2714, European Language Resources Association, Miyazaki, Japan, <https://aclanthology.org/L18-1429>.

Irina NIKOLAEVA (2007), *Finiteness: Theoretical and empirical foundations*, Oxford University Press, Oxford.

Mary PASTER (2016), Alternations: Stems and allomorphy, in Andrew HIPPISEY and Gregory STUMP, editors, *The Cambridge handbook of morphology*, pp. 93–116, Cambridge University Press, Cambridge.

Steven PIANTADOSI, Harry TILY, and Edward GIBSON (2011), Word lengths are optimized for efficient communication, *Proceedings of the National Academy of Sciences*, 108(9):3526–3529, doi:10.1073/pnas.1012551108.

Vito PIRRELLI and Marco BATTISTA (2000), The paradigmatic dimension of stem allomorphy in Italian verb inflection, *Rivista di Linguistica*, 12(2):307–380.

Geoffrey PULLUM and Arnold ZWICKY (1991), A misconceived approach to morphology, *Proceedings of the West Coast Conference on Formal Linguistics*, 10.

R CORE TEAM (2021), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/>.

Robert RATCLIFFE (1998), *The “broken” plural problem in Arabic and comparative Semitic*, Benjamins, Amsterdam.

Robert Henry ROBINS (1959), In defence of WP, *Transactions of the Philological Society*, 58(1):116–144, doi:10.1111/j.1467-968X.1959.tb00301.x.

Robert Henry ROBINS (1997), *A short history of linguistics*, Routledge, New York, 4th edition.

Jerrold SADOCK and Arnold ZWICKY (1985), Speech act distinctions in grammar, in Timothy SHOPEN, editor, *Language typology and syntactic description. Volume 1*, pp. 155–196, Cambridge University Press, Cambridge.

Gerald SANDERS (1988), Zero derivation and the overt analogue criterion, in Michael HAMMOND and Michael NOONAN, editors, *Theoretical Morphology*, pp. 155–175, Academic Press, San Diego, CA.

Gene SCHRAMM (1962), An outline of Classical Arabic verb structure, *Language*, 38(4):360–375, doi:10.2307/410672.

Ilja SERŽANT and George MOROZ (2022), Universal attractors in language evolution provide evidence for the kinds of efficiency pressures involved, *Humanities and Social Sciences Communications*, 9(1):1–9, doi:10.1057/s41599-022-01072-0.

Anna SIEWIERSKA (2010), Person asymmetries in zero expression and grammatical functions, in Franck FLORICIC, editor, *Essais de linguistique generale et de typologie linguistique offerts au professeur Denis Creissels à l'occasion de ses 65 ans*, pp. 425–438, Presses de l'École Normale Supérieure, Paris.

Jennifer SMITH (2005), *Phonological augmentation in prominent positions*, Taylor & Francis, Oxfordshire, doi:10.4324/9780203506394.

Jae Jung SONG (2018), *Linguistic typology*, Oxford University Press, Oxford.

Andrew SPENCER (2012), Identifying stems, *Word Structure*, 5(1):88–108, doi:10.3366/word.2012.0021.

Matthew STAVE, Ludger PASCHEN, François PELLEGRINO, and Frank SEIFART (2021), Optimization of morpheme length: A cross-linguistic assessment of Zipf's and Menzerath's laws, *Linguistics Vanguard*, 7(s3), doi:10.1515/lingvan-2019-0076.

Thomas STOLZ and Nataliya LEVKOVYCH (2019), Absence of material exponence, *Language Typology and Universals*, 72(3):373–400, doi:10.1515/stuf-2019-0015.

Gregory STUMP (2001), *Inflectional morphology: A theory of paradigm structure*, Cambridge University Press, Cambridge.

Gregory STUMP and Rafael FINKEL (2013), *Morphological typology: From word to paradigm*, Cambridge University Press, Cambridge.

John SYLAK-GLASSMAN (2016), The composition and use of the Universal Morphological Feature Schema (UniMorph Schema), <https://unimorph.github.io/doc/unimorph-schema.pdf>.

Anna THORNTON (2012), Reduction and maintenance of overabundance. A case study on Italian verb paradigms, *Word Structure*, 5(2):183–207, doi:10.3366/word.2012.0026.

Jochen TROMMER (2012), Ø-exponence, in Jochen TROMMER, editor, *The morphology and phonology of exponence*, pp. 326–354, Oxford University Press, Oxford.

Aki VEHTARI, Andrew GELMAN, and Jonah GABRY (2017), Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, *Statistics and Computing*, 27(5):1413–1432, doi:10.1007/s11222-016-9696-4.

Laurence WHITE, Silvia BENAVIDES-VARELA, and Katalin MÁDY (2020), Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues?, *Journal of Phonetics*, 81:100982, doi:10.1016/J.WOCN.2020.100982.

Jingting YE (2020), Independent and dependent possessive person forms: Three universals, *Studies in Language*, 44(2):363–406, doi:10.1075/sl.19020.ye.

Moira YIP (1988), Template Morphology and the direction of association, *Natural Language & Linguistic Theory*, 6(4):551–577, doi:10.1007/BF00134493.

Daniel ZEMAN, Joakim NIVRE, Mitchell ABRAMS, Elia ACKERMANN, Noëmi AEPLI, Hamid AGHAEL, Željko AGIĆ, Amir AHMADI, Lars AHRENBERG, Chika Kennedy AJEDE, Gabrielè ALEKSANDRAVIČIŪTĖ, Ika ALFINA, Lene ANTONSEN, Katya APLONOVA, Angelina AQUINO, Carolina ARAGON, Maria Jesus ARANZABE, Bilge Nas ARICAN, Órunn ARNARDÓTTIR, Gashaw ARUTIE, Jessica Naraiswari ARWIDARASTI, Masayuki ASAHARA, Deniz Baran ASLAN, Luma ATEYAH, Furkan ATMACA, Mohammed ATTIA, Aitziber ATUTXA, Liesbeth AUGUSTINUS, Elena BADMAEVA, Keerthana BALASUBRAMANI, Miguel BALLESTEROS, Esha BANERJEE, Sebastian BANK, Verginica BARBU MITITELU, Starkaður BARKARSON, Rodolfo BASILE, Victoria BASMOV, Colin BATCHELOR, John BAUER, Seyyit Talha BEDIR, Kepa BENGOTXEA, Gözde BERK, Yevgeni BERZAK, Irshad Ahmad BHAT, Riyaz Ahmad BHAT, Erica BIAGETTI, Eckhard BICK, Agnè BIELINSKIENĖ, Kristín BJARNADÓTTIR, Rogier BLOKLAND, Victoria BOBICEV, Loïc BOIZOU, Emanuel BORGES VÖLKER, Carl BÖRSTELL, Cristina BOSCO, Gosse BOUMA, Sam BOWMAN, Adriane BOYD, Anouck BRAGGAAR, Kristina BROKAITĖ, Aljoscha BURCHARDT, Marie CANDITO, Bernard CARON, Gauthier CARON, Lauren CASSIDY, Tatiana CAVALCANTI, Gülşen CEBIROĞLU ERYİĞİT, Flavio Massimiliano CECCHINI, Giuseppe G. A. CELANO, Slavomír ČÉPLÖ, Neslihan CESUR, Savas CETIN, Özlem ÇETİNOĞLU, Fabricio CHALUB, Shweta CHAUHAN, Ethan CHI, Taishi CHIKA, Yongseok CHO, Jinho CHOI, Jayeol CHUN, Juyeon CHUNG, Alessandra T. CIGNARELLA, Silvie CINKOVÁ, Aurélie COLLOMB, Çağrı ÇÖLTEKİN, Miriam CONNOR, Marine COURTIN, Mihaela CRISTESCU, Philemon DANIEL, Elizabeth DAVIDSON, Marie-Catherine DE MARNEFFE, Valeria DE PAIVA, Mehmet Oguz DERIN, Elvis DE SOUZA, Arantza DIAZ DE ILARRAZA, Carly DICKERSON, Arawinda DINAKARAMANI, Elisa DI NUOVO, Bamba DIONE, Peter DIRIX, Kaja DOBROVOLJC, Timothy DOZAT, Kira DROGANOVA, Puneet DWIVEDI, Hanne ECKHOFF, Sandra EICHE, Marhaba ELI, Ali ELKAHKY, Binyam EPHREM, Olga ERINA, Tomaž ERJAVEC, Aline ETIENNE, Wograine EVELYN, Sidney FACUNDES, Richárd FARKAS, Jannatul FERDAOUSI, Marília FERNANDA, Hector FERNANDEZ ALCALDE, Jennifer FOSTER, Cláudia FREITAS,

Kazunori FUJITA, Katarína GAJDOŠOVÁ, Daniel GALBRAITH, Marcos GARCIA, Moa GÄRDEFORS, Sebastian GARZA, Fabrício Ferraz GERARDI, Kim GERDES, Filip GINTER, Gustavo GODOY, Iakes GOENAGA, Koldo GOJENOLA, Memduh GÖKIRMAK, Yoav GOLDBERG, Xavier GÓMEZ GUINOVART, Berta GONZÁLEZ SAAVEDRA, Bernadeta GRICIŪTĖ, Matias GRIONI, Loïc GROBOL, Normunds GRŪZTIS, Bruno GUILLAUME, Céline GUILLOT-BARBANCE, Tunga GÜNGÖR, Nizar HABASH, Hinrik HAFSTEINSSON, Jan HAJIČ, Jan HAJIČ JR., Mika HÄMÄLÄINEN, Linh HÀ MỸ, Na-Rae HAN, Muhammad Yudistira HANIFMUTI, Sam HARDWICK, Kim HARRIS, Dag HAUG, Johannes HEINECKE, Oliver HELLWIG, Felix HENNIG, Barbora HLADKÁ, Jaroslava HLAVÁČOVÁ, Florinel HOCIUNG, Petter HOHLE, Eva HUBER, Jena HWANG, Takumi IKEDA, Anton Karl INGASON, Radu ION, Elena IRIMIA, Ǫlájídé ISHOLA, Kaoru ITO, Siratun JANNAT, Tomáš JELÍNEK, Apoorva JHA, Anders JOHANNSEN, Hildur JÓNSDÓTTIR, Fredrik JØRGENSEN, Markus JUUTINEN, Sarveswaran K, Hüner KAŞIKARA, Andre KAASEN, Nadezhda KABAEVA, Sylvain KAHANE, Hiroshi KANAYAMA, Jenna KANERVA, Neslihan KARA, Boris KATZ, Tolga KAYADELEN, Jessica KENNEY, Václava KETTNEROVÁ, Jesse KIRCHNER, Elena KLEMENTIEVA, Elena KLYACHKO, Arne KÖHN, Abdullatif KÖKSAL, Kamil KOPACEWICZ, Timo KORKIAKANGAS, Mehmet KÖSE, Natalia KOTSYBA, Jolanta KOVALEVSKAITĖ, Simon KREK, Parameswari KRISHNAMURTHY, Sandra KÜBLER, Oğuzhan KUYRUKÇU, Asli KUZGUN, Sookyoung KWAK, Veronika LAIPPALA, Lucia LAM, Lorenzo LAMBERTINO, Tatiana LANDO, Septina Dian LARASATI, Alexei LAVRENTIEV, John LEE, Phươg LÊ HỒNG, Alessandro LENCI, Saran LERTPRADIT, Herman LEUNG, Maria LEVINA, Cheuk Ying LI, Josie LI, Keying LI, Yuan LI, KyungTae LIM, Bruna LIMA PADOVANI, Krister LINDÉN, Nikola LJUBEŠIĆ, Olga LOGINOVA, Stefano LUSITO, Andry LUTHFI, Mikko LUUKKO, Olga LYASHEVSKAYA, Teresa LYNN, Vivien MACKETANZ, Menel MAHAMDI, Jean MAILLARD, Aibek MAKAZHANOV, Michael MANDL, Christopher MANNING, Ruli MANURUNG, Büşra MARŞAN, Cătălina MĂRĂNDUC, David MAREČEK, Katrin MARHEINECKE, Héctor MARTÍNEZ ALONSO, Lorena MARTÍN-RODRÍGUEZ, André MARTINS, Jan MAŠEK, Hiroshi MATSUDA, Yuji MATSUMOTO, Alessandro MAZZEI, Ryan McDONALD, Sarah MCGUINNESS, Gustavo MENDONÇA, Tatiana MERZHEVICH, Niko MIEKKA, Karina MISCHENKOVA, Margarita MISIRPASHAYEVA, Anna MISSILÄ, Cătălin MITITELU, Maria MITROFAN, Yusuke MIYAO, AmirHossein MOJIRI FOROUSHANI, Judit MOLNÁR, Amirsaeid MOLOODI, Simonetta MONTEMAGNI, Amir MORE, Laura MORENO ROMERO, Giovanni MORETTI, Keiko Sophie MORI, Shinsuke MORI, Tomohiko MORIOKA, Shigeki MORO, Bjartur MORTENSEN, Bohdan MOSKALEVSKYI, Kadri MUISCHNEK, Robert MUNRO, Yugo MURAWAKI, Kaili MÜÜRISSEP, Pinkey NAINWANI, Mariam NAKHLÉ, Juan Ignacio NAVARRO HORŇIAČEK, Anna NEDOLUZHKO, Gunta NEŠPORE-BĚRZKALNE, Manuela NEVACI, Lươg NGUYỄN THỊ, Huyền NGUYỄN THỊ MINH, Yoshihiro NIKAIIDO, Vitaly

NIKOLAEV, Rattima NITISAROJ, Alireza NOURIAN, Hanna NURMI, Stina OJALA, Atul Kr. OJHA, Adédayo OLÚÒKUN, Mai OMURA, Emeka ONWUEGBUZIA, Petya OSENOVA, Robert ÖSTLING, Lilja ØVRELID, Şaziye Betül ÖZATEŞ, Merve ÖZÇELİK, Arzucan ÖZGÜR, Balkız ÖZTÜRK BAŞARAN, Hyunji Hayley PARK, Niko PARTANEN, Elena PASCUAL, Marco PASSAROTTI, Agnieszka PATEJUK, Guilherme PAULINO-PASSOS, Angelika PELJAK-ŁAPIŃSKA, Siyao PENG, Cenel-Augusto PEREZ, Natalia PERKOVA, Guy PERRIER, Slav PETROV, Daria PETROVA, Jason PHELAN, Jussi PIITULAINEN, Tommi A. PIRINEN, Emily PITLER, Barbara PLANK, Thierry POIBEAU, Larisa PONOMAREVA, Martin POPEL, Lauma PRETKALNIŅA, Sophie PRÉVOST, Prokopis PROKOPIDIS, Adam PRZEPIÓRKOWSKI, Tiina PUOLAKAINEN, Sampo PYYSALO, Peng QI, Andriela RÄÄBIS, Alexandre RADEMAKER, Mizanur RAHOMAN, Taraka RAMA, Loganathan RAMASAMY, Carlos RAMISCH, Fam RASHEL, Mohammad Sadegh RASOOLI, Vinit RAVISHANKAR, Livy REAL, Petru REBEJA, Siva REDDY, Mathilde REGNAULT, Georg REHM, Ivan RIABOV, Michael RIESSLER, Erika RIMKUTĖ, Larissa RINALDI, Laura RITUMA, Putri RIZQIYAH, Luisa ROCHA, Eiríkur RÖGNVALDSSON, Mykhailo ROMANENKO, Rudolf ROSA, Valentin ROŞCA, Davide ROVATI, Olga RUDINA, Jack RUETER, Kristján RÚNARSSON, Shoval SADDE, Pegah SAFARI, Benoît SAGOT, Aleksi SAHALA, Shadi SALEH, Alessio SALOMONI, Tanja SAMARDŽIĆ, Stephanie SAMSON, Manuela SANGUINETTI, Ezgi SANIYAR, Dage SÄRG, Baiba SAULTE, Yanin SAWANAKUNANON, Shefali SAXENA, Kevin SCANNELL, Salvatore SCARLATA, Nathan SCHNEIDER, Sebastian SCHUSTER, Lane SCHWARTZ, Djamé SEDDAH, Wolfgang SEEKER, Mojgan SERAJI, Syeda SHAHZADI, Mo SHEN, Atsuko SHIMADA, Hiroyuki SHIRASU, Yana SHISHKINA, Muh SHOHIBUSSIRRI, Dmitry SICHINAVA, Janine SIEWERT, Einar Freyr SIGURÐSSON, Aline SILVEIRA, Natalia SILVEIRA, Maria SIMI, Radu SIMIONESCU, Katalin SIMKÓ, Mária ŠIMKOVÁ, Kiril SIMOV, Maria SKACHEDUBOVA, Aaron SMITH, Isabela SOARES-BASTOS, Shafi SOUROV, Carolyn SPADINE, Rachele SPRUGNOLI, Steinór STEINGRÍMSSON, Antonio STELLA, Milan STRAKA, Emmett STRICKLAND, Jana STRNADOVÁ, Alane SUHR, Yogi Lesmana SULESTIO, Umut SULUBACAK, Shingo SUZUKI, Zsolt SZÁNTÓ, Chihiro TAGUCHI, Dima TAJI, Yuta TAKAHASHI, Fabio TAMBURINI, Mary Ann C. TAN, Takaaki TANAKA, Dipta TANAYA, Samson TELLA, Isabelle TELLIER, Marinella TESTORI, Guillaume THOMAS, Liisi TORGa, Marsida TOSKA, Trond TROSTERUD, Anna TRUKHINA, Reut TSARFATY, Utku TÜRK, Francis TYERS, Sumire UEMATSU, Roman UNTILOV, Zdeňka UREŠOVÁ, Larraitx URIA, Hans USZKOREIT, Andrius UTKA, Sowmya VAJJALA, Rob VAN DER GOOT, Martine VANHOVE, Daniel VAN NIEKERK, Gertjan VAN NOORD, Viktor VARGA, Eric VILLEMONTÉ DE LA CLERGERIE, Veronika VINCZE, Natalia VLASOVA, Aya WAKASA, Joel C. WALLENBERG, Lars WALLIN, Abigail WALSH, Jing Xian WANG, Jonathan North WASHINGTON, Maximilan WENDT, Paul WIDMER, Sri Hartati WIJONO, Seyi WILLIAMS, Mats WIRÉN, Christian WITTERN, Tsegay

WOLDEMARIAM, Tak-sum WONG, Alina WRÓBLEWSKA, Mary YAKO, Kayo YAMASHITA, Naoki YAMAZAKI, Chunxiao YAN, Koichi YASUOKA, Marat M. YAVRUMYAN, Arife Betül YENICE, Olcay Taner YILDIZ, Zhuoran YU, Arlisa YULIAWATI, Zdeněk ŽABOKRTSKÝ, Shorouq ZAHRA, Amir ZELDES, He ZHOU, Hanzhi ZHU, Anna ZHURAVLEVA, and Rayan ZIANE (2023), Universal dependencies 2.13, <http://hdl.handle.net/11234/1-5287>.

George Kingsley ZIPF (1935), *The psychobiology of language: An introduction to dynamic philology*, MIT Press, Cambridge, MA.

Arnold ZWICKY (1985), How to describe inflection, in Mary NIEPOKUJ, Mary VAN CLAY, Vassiliki NIKIFORIDOU, and Deborah FEDER, editors, *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society*, pp. 372–386, Berkeley Linguistics Society, Berkeley, CA.

Laura Becker

© 0000-0002-1835-9404

Department of Linguistics
University of Freiburg
Belfortstraße 18,
79098 Freiburg im Breisgau, Germany

Laura Becker (2024), Zero marking in inflection: A token-based approach, *Journal of Language Modelling*, 12(2):349–413

doi <https://dx.doi.org/10.15398/jlm.v12i2.361>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

cc  <http://creativecommons.org/licenses/by/4.0/>

An analogical approach to the typology of inflectional complexity

Matías Guzmán Naranjo
University of Freiburg

ABSTRACT

This paper studies the inflectional complexity of nouns, verbs and adjectives in 137 datasets, across 71 languages. I follow Ackerman and Malouf (2013) in distinguishing between E(numerative) complexity and I(ntegrative) complexity. The first one encompasses aspects of inflection, like the number of principal parts, paradigm size, and number of exponents, while the second one captures the implicative relations between paradigm cells (how difficult it is to predict one cell of a paradigm knowing a different cell). I provide a formalism and computational implementation to estimate both I- and E-complexity expressed through Word and Paradigm morphology (Blevins 2006, 2016), which is flexible and powerful enough for typological research. The results show that, as suggested by Ackerman and Malouf (2013), I-complexity is relatively low across the languages in the sample, with only two clear exceptions (Navajo and Yaitepec-Chatino). The results also show that E-complexity can vary considerably cross-linguistically. Finally, I show there is a clear correlation between I- and E-complexity.

Keywords:
inflectional
complexity,
typology,
analogy, Word
and Paradigm
morphology

The study of morphological complexity has a long history in linguistics and typology (see for example Greenberg 1960, for an early approach), and has seen a renewed interest in recent years (Ackerman and Malouf 2013; Cotterell *et al.* 2019; Bentz *et al.* 2022). However, there is very little unity or agreement regarding how we should measure inflectional complexity, and whether the proposed metrics are cross-linguistically comparable (Igartua and Santazilia 2018; Gutierrez-Vasques and Mijangos 2019; Bentz *et al.* 2022, 2016; Arkadiev and Gardani 2020, among many others). Igartua and Santazilia (2018, p. 439) for example, define morphological complexity as “the extent to which formal differences in inflectional paradigms are semantically or phonologically unmotivated” (i.e. the amount of allomorphy in a morphological system). In contrast, Sinnemäki and Di Garbo (2018, p. 8), following Bickel and Nichols (2007) and Bickel and Nichols (2013), define the inflectional complexity of a verb as: “the number of morphological categories expressed per word in a maximally inflected verb form.”

One key development in the study of inflectional complexity came from Ackerman and Malouf (2013), who propose a distinction between two fundamentally different types of inflectional complexity: Enumerative (E) complexity and Integrative (I) complexity. The first is the complexity in morphosyntactic distinctions and the way languages encode them (be it exponents, morphs, morphemes, etc.), while the second one is the difficulty a paradigm poses to speakers in terms of implicative relations. Ackerman and Malouf (2013, 429) provide the following definitions:

The I-complexity of an inflectional system reflects the difficulty that a paradigmatic system poses for language users (rather than lexicographers) in information-theoretic terms. (Ackerman and Malouf 2013, p. 429)

I-complexity measures how predictable the realisation of a lexeme is, given knowledge about one (or more) cells of its paradigm. This type of complexity measures implicational relations in a paradigm,

and it is not directly dependent on paradigm shape (what the actual realisations are, or how many cells a paradigm has). In contrast, E-complexity is defined as:

E-complexity [is given by] the number of exponents, inflectional classes, and principal parts (Ackerman and Malouf 2013, p. 429)

E-complexity has received considerable attention in the literature (Stump and Finkel 2013; Bentz *et al.* 2022; Finkel and Stump 2007; Baerman *et al.* 2015; Dressler 2011), from several different perspectives, including inflection class systems, paradigm size, principal parts and number of morphs or morphemes. Some of this work, however, faces some practical and theoretical challenges (see Section 2).

At the same time, while there are multiple computational proposals for capturing I-complexity (Bonami and Beniamine 2016; Cotterell *et al.* 2019; Guzmán Naranjo 2020; Ackerman and Malouf 2013; Marzi *et al.* 2019), most studies have looked at a relatively small samples (< 100 datasets) and the emphasis has not been on cross-linguistic comparison (although see Cotterell *et al.* 2019). This means that we still do not have a good picture of how I-complexity varies across languages and systems. For example, one still open question is how verb, noun and adjective paradigms compare cross-linguistically for the sake of consistency in terms of I-complexity.

The objectives of this paper are twofold: First and foremost, it presents a medium-scale typological study of morphological complexity from a Word and Paradigm perspective (Blevins 2006, 2016; Matthews 1972). And second, it presents a new technique for measuring morphological complexity and provides an efficient computational implementation of it. I argue that it is both feasible and desirable to work from a W&P perspective when doing cross-linguistic comparisons of inflectional systems. I also show that some fundamental problems in morphological typology can be completely bypassed when approached from a W&P perspective.

The paper is structured as follows: Section 2 gives a brief overview of the main ideas and approaches to morphological complexity, as well as word and paradigm morphology. Section 3 describes the datasets and the methods used in the paper. Section 4 presents the results, and Section 5 concludes.

2

BACKGROUND

This section presents a very brief overview of the main ideas of and trends in the morphological complexity literature from two different perspectives. It also discusses some key differences between morpheme-based and W&P approaches to morphology and argues that beyond theoretical considerations, there are practical reasons why the latter is preferable for doing cross-linguistic studies of the complexity of morphological inflection.

Due to the vast amount of research on the topic of morphological complexity (see for example Baerman *et al.* 2015, 2017; Miestamo *et al.* 2008; Bentz *et al.* 2022, for some overviews and recent takes), a full account of these topics is not feasible within the scope of this article, and I will concentrate on some of the more important works on the topic. Similarly, covering the whole debate between different types of morphological theories is not feasible, and I will only discuss some of the more concrete and practical issues.

2.1

Integrative-complexity

The initial work on I-complexity was approached using information theory, and it focused on measuring the conditional entropy between the cells of the paradigm of a lexeme, often using hand-extracted exponents for each cell (Ackerman and Malouf 2013; Bonami and Beniamine 2016; Blevins 2013; Palancar 2021; Parker and Sims 2020, among many others). More recent papers estimate the conditional entropy of a system using LSTMs¹ (Cotterell *et al.* 2019; Court *et al.* 2022) instead of directly calculating it based on extracted exponents.

I will illustrate I-complexity with two simple toy examples in Tables 1 and 2.² Both examples have three inflection classes with two

¹LSTMs are a type of neural network that performs sequence to sequence predictions. In this context, they are trained to predict fully inflected forms from other fully inflected forms (plus lexeme information). The entropy of the system is calculated on the network itself.

²The elements in each cell are meant to be the suffix (markers) which express the cell content. These are just examples, and the actual cell realisations could be achieved by suffixes, infixes, tones, etc.

and three cells, but the exponent structure is completely different. The system of Table 1, Language 1, only has two markers, *-i* and *-o*, while the system in Table 2, Language 2, contains 9 different markers: *-i*, *-e*, *-a*, *-u*, *-o*, *-ø*, *-ik*, *-ek*, *-æ*. Language 2 has a higher E-complexity both in terms of the number of exponents and paradigm size, however, the situation is reversed for I-complexity.

	Cell 1	Cell 2	Table 1: I-complexity Language 1
class A	-i	-i	
class B	-o	-i	
class C	-o	-o	

	Cell 1	Cell 2	Cell 3	Table 2: I-complexity Language 2
class A	-i	-e	-a	
class B	-u	-o	-ø	
class C	-ik	-ek	-æ	

Following Ackerman and Malouf (2013), we can measure the I-complexity of each system using conditional entropy. The entropy of a cell X, $H(X)$, in a paradigm can be calculated as:

$$(1) \quad H(X) = - \sum_i p(x_i) \log_2(p(x_i))$$

Where $p(x_i)$ can be calculated from the frequency of the exponents for a cell across inflection classes, and where, i ranges over contrastive exponents found in a cell. However, for illustration purposes, this example assumes that all inflection classes have the same number of lexemes, meaning we can let i range over inflection classes. For Language 1, the frequency of *-i* for Cell 1 is 1, and the frequency for *-o* is 2, meaning $p(-i) = 1/3$ and $p(-o) = 2/3$, which gives us $H(\text{Cell 1}) = 1/3 \log_2(1/3) + 2/3 \log_2(2/3) = 0.92$. This is a measure of how much information is required to capture Cell 1 for Language 1.

The conditional entropy of a cell X given knowledge of cell Y, $H(X|Y)$, can be calculated as:

$$(2) \quad H(X|Y) = H(X, Y) - H(Y)$$

$$(3) \quad = \sum_i \sum_j p(x_i, y_j) \log_2(p(x_i|y_j))$$

For Language 1, the conditional entropy $H(\text{Cell 1}|\text{Cell 2} = -i) = 1$, and $H(\text{Cell 1}|\text{Cell 2} = -o) = 0$. Then, the average conditional entropy $H(\text{Cell 1}|\text{Cell 2}) = 2/3$ (since *-i* appears in two inflection classes, while *-o* appears in 1). Knowing that for a lexeme Cell 2 has the realisation *-o* provides complete information about what its realisation in Cell 1 must be, namely *-o*. In contrast, knowing that the exponent for Cell 2 is *-i* does not provide information about the realisation of Cell 1 because a lexeme with *-i* for Cell 2, can either *-i* or *-o* in Cell 1. Because Language 1 has a symmetric structure, $H(\text{Cell 2}|\text{Cell 1})$ is also $2/3$, meaning that the average pairwise³ conditional entropy is $2/3$. The results for Language 2 are very different. In this case, every cell provides complete information about every other cell in the paradigm of a lexeme, which means that for all pairwise conditional entropy calculations the results are 0, and the average conditional entropy of Language 2 is 0. This very simplified example illustrates the fact that the average E-complexity of a language (measured in terms of paradigm size or the number of markers) is not necessarily correlated with its I-complexity.

While using conditional entropy is still a relatively popular method to estimate I-complexity, an alternative approach is based on the accuracy of classification, instead of conditional entropy (Guzmán Naranjo 2020; Bonami and Pellegrini 2022). Instead of measuring the amount of information a cell provides about another cell, one can train a classifier⁴ on the content of one cell of a lexeme to predict the realisation of another cell for that lexeme.

As an example of classification, if we are dealing with nominal inflection, we can train a classifier to predict the accusative singular from the nominative singular. The accuracy obtained by that classifier (under cross-validation) is then a measure of the I-complexity of the paradigm. If a classifier has a perfect accuracy of 1 predicting the

³See Bonami and Beniamine (2016) for a method to calculate the conditional entropy taking multiple cells into account.

⁴Here classifier is understood as any system which takes some word form as an input and assigns it to a class. The method used could be a rule-based system, logistic regression, neural network, etc. For the purposes of modelling inflection, we usually train classifiers on the phonology and semantics of the forms in question, and predict the inflection class from this information.

inflection class of all lexemes in a language, then we can say that there is effectively no I-complexity to that inflection class. The important point is that, just as with conditional entropy, the I-complexity of a system is mostly independent of the number of inflection classes or exponents in an inflectional system. If there is enough information for the classifier to have perfect accuracy, then the I-complexity of the system will be 0.

Using the previous example, the accuracy for Language 2 will be 1 for all cell pairs, because every cell provides complete information about every other cell in the paradigm of a lexeme, which means that the complexity of Language 2 is also 0. For Language 1, the accuracy of predicting (i.e the number of correct predictions over total number of items) Cell 1 from Cell 2 (and the other way around) is $2/3$ (because on average we will be able to correctly predict the realisation 2 out of 3 times). This means that the average complexity of Language 1 is ~ 0.67 .

One advantage of using a predictive technique instead of estimating conditional entropy using LSTMs is that we can easily make use of classifiers that work well even on very small datasets. LSTMs, due to the way they are trained, can struggle with small datasets. Cotterell *et al.* (2019, 336), for example, restrict their study to languages with at least 700 lexemes, because the specific model requires relatively large datasets to achieve acceptable accuracies. As we will see in the results section, these much simpler models perform well on much smaller datasets.

Enumerative complexity

2.2

The initial definition of E-complexity covered the number of principal parts, exponents, and inflection classes. In this section, I will discuss some of the studies that have looked at these, and a few other aspects of E-complexity.

Principal parts

2.2.1

Principal parts are defined as the cells in the paradigm of a lexeme which a speaker needs to know in order to be able to deduce all other cells (Finkel and Stump 2007). For example, it is often proposed that the Latin verb system has 4 principal parts, which a speaker would

need to know in order to be able to produce all other inflected forms of the verb, these are the first person singular present indicative active, active present infinitive, first person singular perfect indicative active, and the passive perfect participle (or future participle) (Bennett 1918). While this is, in principle, a relatively straightforward way of quantifying the complexity of an inflectional system determining the number of principal parts is not straightforward, and will vary depending on the approach one takes to how principal parts should behave within and across paradigms (Finkel and Stump 2007). In this paper, I will not directly consider counting principal parts, but I will come back to the question during the discussion of the results.

2.2.2

Inflection classes

Measuring inflectional complexity in terms of the number of inflection classes is, in theory, straightforward: one simply counts how many inflection classes there are in a system. Although the idea of inflection classes might seem intuitive, the task of counting inflection classes is particularly difficult. Some early work on complexity approached the problem from this perspective (Carstairs 1983; Carstairs-McCarthy 1994), but it has lost favour during the past decade (Sims and Parker 2016). One of the reasons is the move towards questions of I-complexity, but another is that counting inflection classes is anything but simple. For example, Parker and Sims (2020) show how non-trivial it is to count inflection classes for Russian, a very well studied language. A similar conclusion is reached by Beniamine and Guzmán Naranjo (2021), who show that if taken at a surface level, it is difficult to determine the number of inflection classes a language can have (cf. Beniamine Forthcoming).

There are several reasons why counting inflection classes is particularly difficult, but it mainly boils down to the fact that identifying whether two lexemes belong to the same inflection class or not is not easy to operationalize. As a simple example, consider irregular verbs, or partially irregular verbs, or defective verbs. Whatever decision one makes regarding the inflection class they belong to or not, will affect the number of inflection classes.⁵

⁵See Section 2.3 for some further discussions on the challenges of cross-linguistic morphological analysis.

The first approach to examining the complexity in the exponents of an inflectional system comes from Greenberg's work on inflectional complexity (Greenberg 1960). Greenberg proposes a method based on indices of synthesis, agglutination, compounding, derivation, gross-inflection index, prefixation, suffixation, isolation, pure inflection index, and concord. These indices are calculated as the ratio of two formal elements, given their frequencies in a text.⁶ For example, the gross inflection index is the ratio of words to inflectional morphemes in a language corpus (Greenberg 1960, 186). A language in which this ratio is 1 will have one inflectional morpheme per word and thus a very low inflectional complexity, while languages with high inflectional complexity will have ratios much lower than 1. Typological work on different aspects of E-complexity is abundant, I will focus on a few recent examples.

While ideas similar to the inflectional index have remained present in more recent work on inflectional complexity (see below), several recent studies have focused on the number of morphosyntactic distinctions marked through inflection (Lupyan and Dale 2010; Bentz and Winter 2013; Cotterell *et al.* 2019). These studies tend to use typological datasets like the World Atlas of Language Structures (Dryer and Haspelmath 2013) or similar databases. A well-known example is Lupyan and Dale (2010), who use hand-annotated features in WALS like degree of syncretism, the number of morphosyntactic categories expressed by the verb, presence of noun/verb agreement, presence of inflectional evidentiality, presence of inflectional negation, among others, as measures of morphological complexity. The idea is that if a language makes more morphosyntactic distinctions in a paradigm, then it is more complex than a language that makes fewer morphosyntactic distinctions in the same paradigm. A similar approach is also taken by Bentz and Winter (2013) in a more recent study. Effectively, these

⁶Greenberg uses rather short texts of 100 words, which, as he admits, leads to only very preliminary results.

studies use paradigm size as a measure of morphological complexity.⁷

A different set of metrics based on corpora (Gutierrez-Vasques and Mijangos 2018; Oh and Pellegrino 2022) try to estimate exponent complexity indirectly. Perhaps the simplest is the type-token ratio (TTR) (Juola 1998, 2008; Kettunen 2014). The idea behind the TTR is that if there is a 1-to-1 relation between word types and word tokens,⁸ then this means that there is a very high degree of inflection in the language, and thus the language has very high morphological complexity. A TTR closer to 0 indicates lower morphological complexity. In practice, TTR values range between 0.05 and 0.2 or 0.6 (Kettunen 2014).⁹

2.2.4

Other corpus metrics

Another proposed method for measuring morphological complexity is to calculate the perplexity¹⁰ of sublexical units (Gutierrez-Vasques and Mijangos 2018). In a segmented word, one can calculate the conditional entropy or perplexity of the units within a single word. Low conditional entropy means higher predictability, and thus lower morphological complexity. This method relies on morphological segmentations. Gutierrez-Vasques and Mijangos (2018) rely on automatic segmentation produced by Morfessor (Smit *et al.* 2014). Other corpus-based metrics include word entropy¹¹ (Bentz and Alikaniotis 2016), which measures the amount of information carried by a word based

⁷Arguably, some of the features considered in these approaches, like degree of syncretism, is not directly about paradigm size, but rather paradigm structure. However, most other metrics are proxies for paradigm size.

⁸Notice this never happens due to Zipfian effects.

⁹The difference lies in whether one normalises the corpus size or not. Because corpus size can have a sizeable impact on TTR, some authors have suggested taking the moving average of the TTR across a fixed sub-corpus length (Covington and McFall 2008, 2010). Doing this ensures that when comparing the complexity in two different sized corpora, the TTR is measured on sub-corpora of roughly the same size.

¹⁰Perplexity can be related to entropy as: $P = 2^H$, where H is the entropy.

¹¹These entropy metrics measure the distribution of words in a corpus and are not to be confused with other entropy measures like those of Ackerman and Malouf (2013), which measure the distribution of inflectional patterns in a lexicon.

on its probability distribution in a corpus; the relative entropy of word structure (Koplenig *et al.* 2017), which is based on a compression algorithm; and word alignment measure (Bentz *et al.* 2016), which assumes that for languages with morphologically complex words, those will be translated into several independent words in morphologically simpler languages.¹²

Although corpus-based metrics have the advantage of not requiring human decisions, they also have a clear downside: they cannot distinguish inflection from derivation and other morphological processes. All current methods based on corpora conflate morphological complexity arising from derivation and morphological complexity which arises from inflection. Moreover, in most implementations, these methods do not separate the complexity of different subsystems within a language. It is possible for a language to have a very high inflectional complexity in the nominal domain, but a very low inflectional complexity in the verbal domain, or the other way around. While this could be explored with tagged corpora, I am not aware of studies which do this.

Complexity correlations and trade-offs

2.2.5

Despite the proliferation of complexity metrics, Bentz *et al.* (2016) argue that most metrics proposed in typology, either based on corpora or hand annotations, are highly correlated with each other. To do this, the authors propose a method to estimate an aggregated metric of inflectional complexity based on WALS features. The process is as follows. First, the authors identify 28 features that they argue to be indicative of the morphological complexity of a language (e.g. number of genders, number of cases, presence of morphological tense marking, etc.). Then, they normalise the values for each feature to be between 0 and 1 in order to make them comparable. Finally, the authors take the mean value of all 28 features for each language. The authors then estimate the correlations of this complexity index with estimates for several corpus-based complexity indices estimated from Bible translations. The fact that Bentz *et al.* (2016) find a relatively

¹²See also Oh and Pellegrino (2022) for a comparison and evaluation of different corpus-based metrics of morphological complexity.

high correlation between all these metrics is taken by the authors as an indication that they indeed capture the same phenomenon.

Finally, another question that has received some attention regarding complexity is whether there are trade-offs between the local complexity of different domains (morphology and syntax). Several studies have found trade-offs between different types of complexity (Koplenig *et al.* 2017; Oh and Pellegrino 2022; Bentz *et al.* 2022). Some work that has looked at E- and I-complexity has proposed that there are trade-offs between the two (Gutierrez-Vasques and Mijangos 2019; Cotterell *et al.* 2019). Gutierrez-Vasques and Mijangos (2019) use the metrics proposed by Bentz *et al.* (2016) for measuring E-complexity, which are based on aggregating 28 morphological features found in WALS. Cotterell *et al.* (2019) use a simpler metric based on the number of cells in a paradigm.¹³ Generally, these studies have found some sort of trade-off between their definition of E-complexity and I-complexity.

2.3

Word and Paradigm morphology for typology

Although intuitive, approaches based on morpheme or morph segmentations face a challenge: segmenting words is difficult and depends on theory and tradition.¹⁴ The key idea here is that it is not always easy to compare segmentations across languages, and even within languages, linguists face what is called the segmentation problem (Spencer 2012), i.e. how to segment words into sublexical units like stems, morphs or morphemes. That is, it is not just that segmenting words into morphemes is difficult, but it can be a problem without a determined solution. Things can be even more complex if one considers that some theories propose zero morphemes, or very complex and abstract morph sequences. In order to compare the complexity of two languages based on metrics that rely on morph or morpheme segmentations, the principles behind the segmentation decisions need to be consistent for all languages, and application needs to be independent of linguistic tradi-

¹³Recall most E-complexity metrics are correlated with, and a proxy for paradigm size.

¹⁴For the opposite view the reader can look at Manova *et al.* 2020.

tions associated with the languages in question.¹⁵ As far as I am aware, there are no clear formalisation for how this should be resolved for typological comparison.¹⁶

In several of the approaches to E-complexity mentioned in the previous section, segmentation of words into morphs or morphemes plays a crucial role (e.g. mean number of morphemes per word). However, segmentation-based approaches to morphology from a cross-linguistic perspective have issues which are not easy to overcome. The first issue worth discussing is that of the definition, delimitation and identification of morph and morpheme boundaries. This is a problem without a simple solution. This has been noted before with regards to morphological complexity. Greenberg (1960, p. 188) notes that:

Basic to the synthetic index as well as most of the others is the possibility of segmenting any utterance in a language into a definite number of meaningful sequences which cannot be subject to further division. Such a unit is called a morph. There are clearly divisions which are completely justified and which every analyst would make. For example, everyone would divide English *eating* into *eat-ing* and say that there were two units. There are other divisions which are just as clearly unjustified. For example, the analysis of *chair* into *ch-*, “wooden object,” and *-air*, “something to sit on,” would be universally rejected. There is, however, an intermediate area of uncertainty in which opinions differ. Should, for example, English *deceive* be analyzed into *de-* and *-ceive*. (Greenberg 1960, p. 188)

This relates to the segmentation problem (Spencer 2012). The implication of this is that trying to do automatic, or even semi-automatic morpheme identification on large datasets is not feasible.¹⁷ More

¹⁵By this I mean how linguists analyse sublexical units like phonemes, tones, or discontinuous stems (i.e. roots in Semitic), zero morphemes, so-called subtractive morphology, etc.

¹⁶Though see below for a computer-aided approach, as well as Sagot and Walther 2011 and Walther and Sagot 2011 for some early approaches in this direction.

¹⁷While tools like Morfessor can, under some circumstances, do a decent job of approximating human judgements in morpheme segmentation, these are

importantly, segmentation done by linguists is not necessarily objective and will be influenced by different theoretical perspectives, and linguistic traditions (see Bonami and Beniamine 2021 for a discussion on stem segmentation). For example, while it is common to view stems in Semitic languages as discontinuous triconsonantal roots, it is not common to take a similar approach for European languages, instead preferring ideas like stem mutation.

The consequence of these issues for linguistic typology is that cross-linguistic comparison is heavily dependent on the individual decisions made by the individual linguists writing the grammars. Morphological analysis and segmentation is usually taken as a given, and it is not possible to be certain that the guiding principles for morpheme segmentation are consistent across languages.

Although there are some attempts at computational formalisations of morpheme-based approaches (Rathi *et al.* 2022), I am not aware of large-scale validations of these for the purpose of studying inflectional morphology cross-linguistically.

The alternative approach is to take whole, fully inflected words and their relations in a paradigm as a starting point of linguistic comparison. If we define a systematic approach to finding relations between fully inflected words (see the next section), then we can be sure that all languages in our sample are analysed using the exact same principles. If we focus on fully inflected words, the issues related to segmentation and morph(eme) boundaries disappear.

Perhaps the main counterargument one can leverage against W&P morphology is that one needs to provide a solid, cross-linguistically valid, definition of what a word is. It has been argued that such a task is impossible (Haspelmath 2011), and Greenberg himself points out the issues with defining word units (Greenberg 1960). While it is true that identifying words can be challenging, it must be noted that this is also a necessary step in all morpheme-based approaches to inflectional complexity I am aware of. In order to estimate metrics related to paradigm size, one first needs to decide which elements belong to a paradigm and which elements do not. This requires at least a definition of words. If one wants to count whether negation is expressed through

nowhere near good enough for the task at hand, and the quality of the output greatly varies with the type of input provided.

inflection or not, one needs to distinguish between what constitutes one or two words. If one wants to calculate the number of morphemes to words ratio one needs to delimit words. And so on. Even the corpus-based metrics discussed in the previous section require orthographic word segmentation to work.

The difficulty of delimiting words, and having a systematic, cross-linguistically valid definition of what a word is, is not an argument in favour of morpheme-based approaches to inflectional complexity, nor is it an argument against W&P approaches. My solution in this paper is the same as with many typological studies: I trust the grammars (or in this case the datasets). Even if different languages require different criteria for defining and delimiting words, I will assume that the authors of the resources I use (see next section) applied the correct and relevant criteria consistently for each language in question.¹⁸

MATERIALS AND METHODS

3

Datasets

3.1

For this study,¹⁹ I mostly rely on Unimorph data which was available in January 2021 (Kirov *et al.* 2018).²⁰ Additionally, I include the following datasets:

¹⁸The only technique that I am aware of, which can completely ignore the issue of words is based on compression algorithms (Moscoso del Prado 2011; Ehret 2021). This type of complexity is also known as Kolmogorov Complexity. These calculate the compression rate of a corpus for a given language (how much a compression algorithm can compress a corpus), and compare that result with the compression rate of either another language, or a modified (e.g. lemmatized) version of the same corpus. For reasons of space, I will not discuss this approach here.

¹⁹All datasets and code can be found at <https://doi.org/10.5281/zenodo.11147171>.

²⁰I am aware that Unimorph has included some additional datasets since then, but our approach is computationally too intensive for us to keep adding languages indefinitely. With my dataset, it took me around 6 months to process all paradigms.

- Russian nouns (Guzmán Naranjo 2020)
- Kasem nouns (Guzmán Naranjo 2019a)
- Latvian nouns (Beniamine and Guzmán Naranjo 2021)
- Hungarian nouns (Beniamine and Guzmán Naranjo 2021)
- French verbs (Bonami *et al.* 2014)
- Arabic nouns (Beniamine 2018)
- Portuguese verbs (Beniamine *et al.* 2021)
- English verbs (CELEX, Baayen *et al.* 1996)
- Latin nouns (Pellegrini and Passarotti 2018)
- Latin verbs (Pellegrini and Passarotti 2018)
- Navajo verbs (Beniamine 2018)
- Yaitepec verbs (Feist and Palancar 2015)
- Zenzotepec verbs (Feist and Palancar 2015)

In total, this makes for 137 datasets across 71 languages for nouns, adjectives, and verbs. The size of these datasets vary considerably, from some languages having a few hundred lexemes, to others containing over 40,000 lexemes. To be able to better compare results, I created random subsamples of 200, 500, 1000, 2000, and 5000 lexemes for each dataset (when available). Although I am aware of some issues with the Unimorph datasets,²¹ I only performed minimal hand corrections. These datasets are structured in long format with three main columns: lexeme, cell, inflected form. Table 3 shows an example of this structure for the Spanish verb *cantar* ‘sing’. All datasets are in orthographic form, except for those listed above, which were converted to a phonemic representation. No other information is required or provided in these datasets.

A final note about the data is that I included all cells listed in unimorph, including elements separated by spaces. These can be inflected forms with pronouns/clitics (like in Romance), single words made up of two elements but which inflect like a single lexeme (like *high school*), or periphrasis. About 25% of the datasets contain at least one form that

²¹The Hungarian and Latvian dataset are effectively hand-corrected unimorph datasets, for which Beniamine and Guzmán Naranjo (2021) remove multiple mistakes present in the original data. Similarly, the Arabic nouns dataset was hand-corrected by Beniamine (2018).

Lexeme	Cell	Inflected form	Table 3: Example of basic data structure
cantar	1.sg.pres.ind	canto	
cantar	2.sg.pres.ind	cantas	
cantar	3.sg.pres.ind	canta	
...	

fits this description. For most purposes periphrastic forms behave almost exactly as non periphrastic ones and do not have an impact on the analysis. While it would be possible to exclude all forms containing spaces, leaving them in for the analysis ensures that we are not arbitrarily reducing the complexity of any of the systems in question.

Methods

3.2

In order to estimate the complexity of a morphological system we need a formal model of that system, and from this formal model, we can then estimate the I- and E-complexity of the system. Word-based models of morphology can be divided into two main camps: symbolic and non-symbolic. Under non-symbolic models there are approaches like LSTMs (Cotterell *et al.* 2019; Malouf 2017; Elsner *et al.* 2022; Cardillo *et al.* 2018) or linear discriminative learning (Baayen *et al.* 2019a,b). Non-symbolic models do not require any type of explicit morphological structures, and can predict one cell in the paradigm of a lexeme from another cell or from a meaning without any sort of symbolic manipulating of the strings (see also Elsner *et al.* 2019, for a recent overview). In these types of approaches there are no explicit representations of sublexical units above the grapheme level, instead, they treat words as sequences of individual letters and the cell in the paradigm they realise. LSTMs are trained to predict sequences from sequences. In the case of morphological inflection, they can predict one inflected form from another directly or from its lexeme meaning and cell in the paradigm (depending on the setup).²² In non-symbolic approaches, there are no explicit representations of proportions in the style $Xa \rightleftharpoons Xb$. Despite their impressive performance,

²²See for example Cotterell *et al.* 2019 or Malouf 2017 for more in-detail descriptions of how LSTMs work for morphological reinflection tasks.

non-symbolic models are not appropriate to our objectives for two main reasons. First, existing implementations are too slow to be applicable to large datasets with many languages, or even to languages with many cells. Second, while it is possible to use these systems to estimate I-complexity, I am unaware of any method for estimating E-complexity from the models themselves. Studies that have used LSTMs to explore morphological complexity (Cotterell *et al.* 2019; Marzi *et al.* 2019; Marzi 2020) have explicitly relied on traditional metrics like the number of paradigm cells.

In contrast, symbolic models use explicit representations of the relations between cells. A symbolic model must be comprised of two independent elements: (1) a system of proportions that express the relations between cells, and (2) a method for assigning a lexeme to the correct proportion. Here, (2) is essentially what has been called the classification problem (Guzmán Naranjo 2020), that is, how to determine the inflection class of an inflected form based on its phonology and semantics. There are multiple proposals for solving (1)²³ and (2).²⁴ In this paper I present a new solution for (1), and, for performance reasons, take a very simple approach to (2). These are described in Section 3.2.1 and 3.2.2, respectively.

3.2.1

Analogical proportions

At the core of symbolic W&P approaches to inflection are the analogical proportions between fully inflected forms. Traditionally, these have been expressed informally as $Xa \Leftarrow X_o$ (sometimes written as $Xa::X_o$, or some variant thereof), where variables are expressed with upper case letters like X or Y, and segments with lower case letters. This proportion expresses the formal relation between two cells in the paradigm of a lexeme. This example would cover alternations like the following: *ata::ato* (X = at), *para::paro* (X = par), etc. However, this notation is not well formalised in the sense that it does not readily work

²³See for example Lepage 1998; Stroppa and Yvon 2005; Federici *et al.* 1995a,b; Carstairs 1998, 1990; Albright and Hayes 1999; Albright *et al.* 2001; Beniamine 2017; Lindsay-Smith *et al.* 2024.

²⁴Among others Bybee and Slobin 1982; Guzmán Naranjo 2019a; Albright and Hayes 1999; Albright *et al.* 2001; Arndt-Lappe 2011, 2014; Eddington 2000; Matthews 2005, 2010, 2013; Skousen 1989; Skousen *et al.* 2002; Skousen 1992.

in a computational implementation. The reason is that it is not precise enough to disambiguate cases where there is more than one variable. For example, the alternation $XaY \rightleftharpoons XYo$ is ambiguous for a (toy example) form like *badan* because it is compatible with either *badan::badno* ($X = \text{bad}$, $Y = \text{n}$) and *badan::bdano* ($X = \text{b}$, $Y = \text{dan}$). The reason is that there are no restrictions on how many segments each variable can match, and there is no way of specifying which of the two *-a-* segments should be matched by the infix.

Computationally implemented formalisms of proportional analogies go back several decades and have taken the form of automata (Lepage 1998, 2004; Stroppa and Yvon 2005; Federici *et al.* 1995a,b; Federici and Pirrelli 1997), string unification (Carstairs 1998, 1990), and more recently context rich alternation patterns (Albright and Hayes 1999; Albright *et al.* 2001; Beniamine 2017), and typed-feature structures (Guzmán Naranjo 2019a). Of these, the only formalisation which would be useful for us given the current state of development and tools for automatic induction is that of Beniamine (2017). The idea of context-rich alternation proportions is that they express alternations in the same spirit of the *X*-notation, but they are stricter, and less flexible, thus producing unambiguous proportions. The general form is $X \rightleftharpoons Y/Z$, meaning that *X* alternates with *Y* in the context of *Z*. For the previous example, the contextual pattern could be written as $a_ \rightleftharpoons _o / \text{bad}_ _$, where the underscores can match single segments, and which would only allow for the match *badan::badno*. While the context-rich pattern approach is certainly an improvement over previous formalisms, it lacks some expressive power and it cannot easily capture more abstract patterns. For example, because this technique does not have anything like named variables, it is not possible to express alternations that rely on reordering (e.g. metathesis) or repeating segments (e.g. reduplication, lengthening). In the formalism by Beniamine (2017), it is not possible to express that a matched segment has to be repeated, or changed to a different position in a string.

In this paper, I propose a modification of context-rich proportions. One key insight of the approach by Beniamine (2017) is that alternations are bounded by one of the edges of the word. While his proposed formalism usually needs to specify a lot of concrete (in terms of specific segments) contextual information, most of the time, all that is

actually needed to avoid ambiguities is to know where from either the right or the left the alternation is taking place. For example, if we have the three pairs: *badan::badno*, *tar::tro* and *kariaban::kariabno* it becomes clear that the alternation targets the vowel between the last two consonants, and that everything before it stays constant.²⁵ It is not actually necessary to specify which consonants are at play, just their positions. Doing so carries an important advantage, namely that we can write more abstract patterns involving any two consonants.

It is important to note here that one of the reasons for using contextual information for Beniamine (2017) is that the context helps disambiguate inflection classes, for example, the context might indicate that the alternation between /a/ and /o/ for some cell pair only happens if the preceding consonant is /n/. This is not important for the present technique because I approach classification as a separate problem which can be solved on its own.

In order to be able to express patterns like metathesis and reduplication, I will rely on named variables. It is important to capture these types of patterns because otherwise the system will need many more individual proportions. For example, in a metathesis situation where the last two segments undergo metathesis: $Xab \rightleftharpoons Xba$, if the system cannot capture this pattern abstractly, we would need specific proportions for every combination of segments that appears across all forms. The same applies to reduplication but see below. The basic notation has the following form:

$$(1) \quad [\langle X1, 2 \rangle a \langle X2, 2 \rangle \rightleftharpoons \langle X1, 2 \rangle o \langle X2, 2 \rangle]$$

Where variables, expressed in angled brackets, are tuples of unique identifiers ($X1, X2, X3, \dots$) and a matching potential, i.e., the number of segments they must match. The matching potential, when expressed with a number, means that the variable must match exactly that many segments. Non-variables are expressed simply as lowercase letters, and \rightleftharpoons separates the two parts of the proportion.

To express that some variables can match arbitrarily many segments, we allow for one named variable in the proportion to use ‘+’ (as in a regular expression) indicating that it can match 1 or more

²⁵One could, of course, characterise this example in terms of syllables, but in this paper I will work exclusively on surface strings due to constraints my data.

segments. However, in order to constrain proportions to specific relative positions within inflected words, proportions need to follow two constraints: (i) in any given pattern, all variables must explicitly state their matching potential (i.e. how many segments they must match), and (ii) only one variable can match arbitrarily many segments. The example in (2) shows what this looks like:

$$(2) \quad [<X1, + > a <X2, 2> \rightleftharpoons <X1, + > o <X2, 2>]$$

The main reason for the restriction on the number of variables which can have + as matching potential is computational. If a proportion contains more than 1 variable with a +, then the proportion can become ambiguous, just like proportions of the form $XaY \rightleftharpoons XYo$ are ambiguous in some cases, like in the case of *badan::badno* and *badan::bdano*. Recall this pattern is ambiguous because it is not clear which *a* should be matched. Fundamentally, any pattern of the form $[<X1, + > <X2, + > \dots]$ will be ambiguous because given a 3 segment string $<abc>$, there is no way to know whether $X1$ should match 1 or 2 segments, and both matches $X1 = <a>$ and $X1 = <ab>$ will be valid. Restricting + to apply to maximally one variable removes the potential for ambiguity. This should be emphasised: the main motivation for this constraint is purely computational: to remove potentially ambiguous proportions. Ambiguous proportions lead to mis-inflection, and would defeat the purpose of the system. From a theoretical perspective, this restriction seems to match our expectations for most inflectional systems. Languages in the dataset do not allow for infixation operations in free positions within words, which is what XaY states. To my knowledge, operations are either constrained to an edge (or distance to it), or apply across the whole word systematically (e.g. harmony).

A potential type of counter example would be a language in which morphological alternations are constrained by lexically-specified phonological or prosodic cues, which can occur anywhere within the word, and which are independent of word boundaries. For example, a language in which stressed syllables undergo an alternation as: *'pokolo::'pakolo*, *po'kolo::po'kalo*, *poko'lo::poko'la*, would require patterns with two variables with + as matching potential. In such a case, the proportions would not be ambiguous because the cue would only

allow one match.²⁶ A second type of potential exception are languages which have been described as having free morph order like Chintang (Bickel *et al.* 2007) and Mari (Luutonen 1997, as cited by Bonami and Crysmann 2013). So far, it remains unclear how these languages should be handled from a W&P perspective. As far as I can tell, none of the languages in my sample require these type of proportions with multiple variables with + matching potential.

Throughout this paper I will refer to these proportions as *local inflection classes*, and contrast it with *global inflection classes*. While two lexemes can share local inflection classes for some set of cell pairs, they do not have to share the same global inflection class. I favour the term *local inflection class* over something more traditional like cell realisation, because these proportions are meaningless for individual cells, and only really express the relation between two cells. It is important to note that a pattern like that in (2) fully determines the relation between the two cells in question (here Cell 1 and Cell 2). If we know the realisation of Cell 1 for some lexeme L, we can unambiguously deduce Cell 2, and the other way around, provided that we know the local inflection class for Cell 1 – Cell 2 in L. If we know one cell of the paradigm of a lexeme, we can deduce all other forms in its paradigm if we know all its local inflection classes (i.e. all proportions to all other cells). Effectively, being able to infer the whole paradigm of a lexeme boils down to the classification problem (i.e. how to determine the inflection class of an inflected form based on its phonology).

At some points in this text, I will use ‘.’ as shorthand for any segment or number of segments: [$\langle X1, + \rangle \rightleftharpoons \langle X1, + \rangle$.], in cases referring to abstract proportions and not concrete analogies. Unlike context-rich proportions, these proportions do not need to contain contextual information. For example, (3) is a proportion which includes contextual information of where a change happens. In this example, ‘c’ acts as context because it is part of the non-contrastive material (i.e. is present in both cells in the same position), and could be subsumed by the variable.

²⁶ Notice this is not the case when stress is not free to wander across the whole word, but is fixed to some position from an edge, like Spanish; or cases in which stress triggers phonological alternations without morphological contrast like in Russian.

$$(3) \quad [\langle X1, + \rangle \text{ c a } \langle X2, 2 \rangle \rightleftharpoons \langle X1, + \rangle \text{ c o } \langle X2, 2 \rangle]$$

Proportions like (3) are unnecessary since the more general pattern (2) already matches this same alternations, and even more cases.

This formalism allows for the following inflectional proportions:

- suffixes [$\langle X1, + \rangle \rightleftharpoons \langle X1, + \rangle .$], [$\langle X1, + \rangle . \rightleftharpoons \langle X1, + \rangle .$],
[$\langle X1, + \rangle . \rightleftharpoons \langle X1, + \rangle$]
- prefixes [$\langle X1, + \rangle \rightleftharpoons . \langle X1, + \rangle$], [$. \langle X1, + \rangle \rightleftharpoons . \langle X1, + \rangle$]
- circumfixes [$\langle X1, + \rangle \rightleftharpoons . \langle X1, + \rangle .$]
- metathesis [$\langle X1, + \rangle \langle X2, 1 \rangle \langle X3, 1 \rangle \rightleftharpoons \langle X1, + \rangle \langle X3, 1 \rangle \langle X2, 1 \rangle$]²⁷
- fixed suprasegmentals (e.g. tones marked with numbers):
[$\langle X1, + \rangle 1 2 \rightleftharpoons \langle X1, + \rangle 3 1$] or [$\langle X1, + \rangle ' \langle X2, 1 \rangle \rightleftharpoons \langle X1, + \rangle \langle X2, 1 \rangle '$]
- reduplication [$\langle X1, + \rangle \langle X2, 1 \rangle \langle X2, 1 \rangle \rightleftharpoons \langle X1, + \rangle \langle X2, 1 \rangle$]
- any combinations of the previous proportions

Except for reduplication, I implemented automatic induction techniques for these proportions (including any and all combinations between suffixes, prefixes, infixes, fixed suprasegmentals and metathesis). That is, the computational implementation can automatically induce proportions required to capture an inflectional system. This induction technique tries to find the most economical, and the fewest proportions that can express the relations between all pairs of cells in a dataset.²⁸ While expressing reduplication in this formalism is straightforward, induction is not. For this paper, I do not implement the induction of reduplication, mostly because it is not very common in in

²⁷ Something to point out regarding metathesis is with the current formalism each pattern has a fixed length, and different length metathesis would require different patterns. For example, carabo:caraoab and carator:caraoort would require two different patterns.

²⁸ For reasons of space I do not discuss the techniques in detail here, but these are provided by the packages *analogyr* (<https://gitlab.com/mguzmann89/analogyR>) and *paradigma* (<https://gitlab.com/mguzmann89/paradigma>).

the inflectional systems of my dataset²⁹ and it is too costly for the induction phase.

This formalization is not without drawbacks. I cannot currently capture patterns that require feature structure representations, like more complex supra-segmental structures or voicing alternations, but extending the system to be able to capture these is straightforward. There is nothing special about feature structure representations, and they could be integrated into the formalism without any changes to how proportions are expressed.³⁰ There are two reasons for why I will work with segments in this paper. The main one is that the datasets do not have feature structure representations, and trying to induce phonological representations from orthography is prone to mistakes, without any guarantees that the resulting representation is any better than the orthographic representation. The second reason is that inference of complex feature alternations like downstep or harmony patterns, is much too complex to be viable for this study.

Similarly, this system cannot represent abstractions which are present in some languages, like reference to morphological structure (German *vorspringen-vor-ge-sprungen* ('jump forward'), where the <ge> occurs between a separable prefix and main verb), or the already mentioned harmony, and voicing alternations. While it would be preferable to have a system which can capture all abstractions of the inflectional system of any and all languages, this is not the aim of this paper. For this paper, we need a system that is capable of producing an inflected cell given another inflected cell in the paradigm of a lexeme. The present formalism is in fact capable of doing this exactly in all cases, even if some of the induced proportions are clumsy or too specific from a human perspective.

²⁹ Arguably, reduplication is the most frequent form of morphology since it is present in languages without affixation. However, it does not play a significant role in our data.

³⁰ The simplest approach would be to allow vectors of phonological features instead of or in addition to the individual segments, this would allow feature structure alternations, for example, given a phonological representation of segments with 2 features (e.g. *high* and *back*, etc.), one could express: [$\langle X1, + \rangle 11 \langle X1, 1 \rangle 1 \Rightarrow \langle X1, + \rangle 10 \langle X1, 1 \rangle 0$]. But other alternatives are possible, like including syllable structure with onsets, nucleus and codas, etc.

Additionally, while patterns like harmony are not directly captured by the system, it is unclear that we need to. For example, in Hungarian, stating that there is an abstract marker -Vk for first singular, and the -V- harmonises with the stem: *lát-látok* ('see') vs. *szeret-szeretek* ('love'), has the same effect as stating that there are two different markers -ek and -ok, with inflection class restrictions. Since capturing the correct classification of such cases is completely straightforward, it is not evident that modelling systems like Hungarian without a specific harmony mechanism should produce different results in terms of estimating the complexity of the system.

I will not discuss induction in detail in this paper, but the following gives a short overview of how induction works. For every dataset:

1. Extract all cell pairs
2. For each cell pair Cell_1:Cell_2, calculate the analogical proportions Cell_1 \rightarrow Cell_2 and the proportion Cell_2 \rightarrow Cell_1 (i.e. the relations as above)
3. Since for each pair of forms there often are several alternative valid proportions:³¹
 - calculate all 'best' proportions
 - after calculating all proportions for all items in Cell_1::Cell_2, rank them by frequency
 - for each form pair keep the most frequent proportion which can apply to it

The result is a system of proportions that fully captures the pairwise relations in the paradigm.

The final issue is how to measure the E-complexity of a paradigm using this system. I will use *fragmentation* as a metric of the relative complexity of a pattern. The fragmentation of a pattern is simply the number of positions with contrastive material between the left and right-hand sides of the proportion, i.e. the number of non-variables. For example, if a pattern like [$<X1, + > \rightleftharpoons$

³¹ Strictly speaking there is not need to choose between the many different proportions, since all induced proportions work correctly for the specific lexeme. This filter is useful for the classification step.

<X1, + > .] (e.g. *sing::sings*) has a fragmentation of 1, while a pattern like [$\langle X1, + \rangle \rightleftharpoons . \langle X1, + \rangle .]$ (e.g. *lachen::gelacht* ‘laugh’ inf::participle) has a fragmentation of 3. This metric is independent of the length and complexity of the actual markers, and their position. Prefixes, infixes and suffix contribute 1 to the the total fragmentation of a pattern. There is a relation between a traditional morph count approach and fragmentation in many situations. In the simplest relation between two cells, syncretism, the fragmentation of the pattern will be 0. If the relation between both cells is that of exclusively affixes or prefixes, then the fragmentation will be 2. A fragmentation of more than 2 means that there are discontinuous inflectional markers, or a prefix-suffix combination.³²

There are several advantages to this technique for measuring E-complexity. First, it completely sidesteps the issue of segmentation. This approach does not need to find morphemes, morphs, stems or any other theoretically motivated sub-lexical unit other than the contrasts between two inflected forms. As a consequence, there is no need to find any sort of optimal multiple alignment of a paradigm, all that is needed are optimal pairwise alignments between cells.

The second advantage is that this method works with relatively small datasets of a few dozen inflected lexemes, at least compared to the types of datasets needed when working with automated morpheme segmentation software, or corpus-based methods. In this approach, we only need paradigms of the lexemes we are interested in, there is no need for large corpora, as is the case with other tools.³³ While in this paper I have tried to include inflectional paradigms as complete as possible, fragmentation could be calculated for just two cells. So even if one has only very sparse, and incomplete information on some

³²This metric is inspired by Bonami and Beniamine (2021), however, in their paper, the definition of the fragmentation of the stem would be equivalent to the number of variables in our proportions, while in this paper I count the number of non-variables in the proportions. In practice, there is very little difference taking one or the other, and additional E-complexity metrics could be developed following similar principles.

³³While tools like Morfessor can be used on similarly small datasets, they will produce better results if trained on larger datasets.

inflectional system, it should be possible to use fragmentation as a measure of its E-complexity.

There are two final caveats regarding fragmentation. The first is that it should be understood as an upper limit of E-complexity. Because the induction method is not perfect, because the data lacks feature structure representation, and because the formalism cannot deal with all types of inflectional patterns found in the languages in question, many of the resulting proportions are more complex than theoretically required. The effect is that the measured fragmentation can be higher than the real fragmentation of the language.

The second one is that fragmentation, and the way it is implemented, assumes that all segment alignments matter, even those that might not correspond to traditionally identified inflectional markers. As an example under the current system, the alternation between the Spanish first person singular and third person singular form in any aspect tense combination will contain an infix: *canto::cantamos* produce $s [<X1, + > <X2, 1 > \rightleftharpoons <X1, + > a m <X2, 1 > s]$ because the *o* is not contrastive material. While there might be arguments against this type of full alignment,³⁴ there two in favor. First, it is unclear from a Word and Paradigm perspective why one should allow some but not all segments to align, especially from a crosslinguistic perspective. Second, implementing an algorithm and computational system to produce alignments which match linguistic intuition is remarkably difficult.

Analogical classification

3.2.2

As mentioned in Section 2, I take a classification-based approach to measure I-complexity. Instead of measuring the entropy of the system, we try to predict the local inflection class of each lexeme based on its phonological properties.³⁵ Complexity of the system is then measured in terms of accuracy. That is, if we can successfully predict all local

³⁴Notice that this is not a unique effect of making pairwise comparison. The same type of alignment would arise in a multiple alignment.

³⁵There is good evidence that other factors like semantics can also play a role in helping predict the inflection class of a lexeme. However, there is no semantic information for most datasets in our sample. For this reason, I will only focus on phonology.

inflection classes of all lexemes in a morphological system, then the accuracy is 1 and the complexity 0. If we can predict none of the proportions the accuracy is 0 and the complexity is 1.

There are many approaches to analogical classification that have been proposed in the literature, including Skousen's Analogical Modelling framework (Skousen 1989; Skousen *et al.* 2002; Skousen 1992; Arndt-Lappe 2011, 2014), TiMBL (Daelemans and Van den Bosch 2005; Daelemans *et al.* 1998), Neutral Networks (Guzmán Naranjo 2019a; Matthews 2005, 2010, 2013), Boosting Trees (Guzmán Naranjo and Bonami 2021; Bonami and Pellegrini 2022), and Minimal Generalization Learner (Albright and Hayes 1999; Albright *et al.* 2001), among others. While most of these techniques would likely perform very well on our data (see below for a comparison), they are too slow in most contexts and do not scale very well.³⁶ Additionally, some authors who have pioneered the use of methods like LSTMs (Cotterell *et al.* 2019) suggests very small datasets (< 500 lexemes) might not be adequate for some of these techniques. Since we are predicting all cells in a paradigm from all other cells pairwise, we need a method that can be trained and cross-validated in as little time as possible, but at the same time is as accurate as possible.³⁷

Here, for reasons of computational efficiency and conceptual simplicity, I will use a *k*-Nearest Neighbours (*k*-NN) algorithm based on an edge-weighted Levenshtein distance. The *k*-NN assigns the local inflection class of a word form based on its phonological similarity to its nearest 5 neighbours.³⁸

³⁶Here 'too slow' should be understood as too slow for most researchers' resources. Of course, with unlimited computing power and enough state of the art GPUs, one could fit as many neural network models as needed within some reasonable time limit. However, most researchers (including the author) working on these issue have finite and limited computing power.

³⁷To give a simple example, the dataset for Latin verbs contains 254 cells in total. This means 64,262 models (from every cell to every other possible cell), and that times 10 to account for cross-validation gives 642,620. Assuming 1 minute to train each model (which is rather optimistic for a Neural Network or Boosting Tree), it would take over a year to capture verbal inflection in Latin.

³⁸I arrived at this number as a good choice for *N* through some previous testing. While it is possible that some systems would be better captured with a different choice for *N*, trying to optimize each dataset would take too long.

		a	s	a	c
	0	1	2	3	4
a	1	0	1	2	3
s	2	1	0	1	2
a	3	2	1	0	1
b	4	3	2	1	1

Table 4:

Levenshtein distance between *casa* and *basa*

The traditional Levenshtein distance (Levenshtein 1966) calculates the minimum number of operations of insertion, deletion, and substitution needed to convert a string *s* into a different string *t*. Table 4 shows the calculation for the strings *casa* and *basa*.³⁹ In this case, the number of operations necessary to transform *casa* into *basa* is one, namely a substitution of *c* for *b*.⁴⁰ Table 4 shows all possible ways of turning *casa* into *basa* using insertions, deletions and substitutions. An operation is represented as a movement on the matrix. Horizontal movement represents deletion, vertical movement represents insertion, and diagonal movement represents either no operation (when there is no change) or substitution. Each operation has a cost, and the values are the accumulated cost. The smallest number of operations is given on the bottom right corner.

While the Levenshtein distance captures the differences between two strings, it ignores where in the strings these differences take place. However, if we want to emphasise that differences at some edges are more important than differences in the middle of the word, then we need an edge-sensitive metric. We use an edge-sensitive metric instead of a symmetric one for two reasons. First, an edge-sensitive metric will give greater weight to what would traditionally be the segments belonging to either suffixes or prefixes, which have been shown to play a greater role in class assignment than segments that belong to what would be analysed as the stem (Guzmán Naranjo 2020). Second, there is ample research showing that the edges of a word play a greater role in class assignment than the inner segments (Guzmán Naranjo 2019b; Arndt-Lappe 2011; Albright *et al.* 2001).

³⁹Here I present the reversed strings. This is for clarity in the following examples below.

⁴⁰It is possible to assign different costs to each operation, but I use a cost of 1 for each for illustration purposes.

Table 5:
Edge weighted Levenshtein
distance between *casa*
and *basa*

	ind		a	s	a	c
ind		0	1	2	3	4
	0	0	1	1.5	1.83	2.08
a	1	1	0	0.5	0.83	1.08
s	2	1.5	1	0	0.33	0.58
a	3	1.83	0.83	0.33	0	0.25
b	4	2.08	1.08	0.58	0.25	0.25

Building an edge-weighted version of the Levenshtein distance is straightforward: we divide the cost of the operation by its relative position to the edge of the word (column ind in Table 5). For example, for the previous pair of *casa* and *basa*, there is one difference in the fourth position from the right edge of the word, meaning that the distance is 0.25. Table 5 shows the corresponding table of operations with accumulated cost. The row and column labelled ‘ind’ show the position of the segment in question from the relevant edge of the word. Notice it is possible to calculate either a right-edge weighted, left-edge weighted, or right-left-edge weighted Levenshtein distance; for the latter we simply average both left-edge and right-edge weighted distances.⁴¹

While most previous approaches to classification have used some form of segmentation into stems and affixes, we can use fully inflected forms as the basis for prediction. The target predictions are the local inflection classes (i.e. the proportions induced as described in the previous section). A simple example will illustrate this. Table 6 shows two cells of the paradigm of 4 Spanish verbs across two inflection classes.

The first step is to build the proportions (already given in the table). These are the local inflection classes we want to predict.

Since we want to measure the I-complexity of the system, we try to predict one cell from the other. Suppose we start with the prediction

⁴¹The reader might find this approach to measuring the distance between words unintuitive. The choice for this metric in this paper was purely practical, and based on initial experiments on a smaller, different datasets, in which it outperformed other Levenshtein-based metrics. It is of course possible that there might be other, better metrics for individual languages, but we were unable to find a better metric that worked consistently better cross-linguistically. See below for some tests.

Gloss	1.SG.PRES.IND	2.SG.PRES.IND	Proportion
touch	toco	tocas	$\langle X1, + \rangle o \rightleftharpoons \langle X1, + \rangle as$
eat	como	comes	$\langle X1, + \rangle o \rightleftharpoons \langle X1, + \rangle es$
sweep	barro	barres	$\langle X1, + \rangle o \rightleftharpoons \langle X1, + \rangle es$
drink	tomo	tomas	$\langle X1, + \rangle o \rightleftharpoons \langle X1, + \rangle as$

Table 6:
Spanish
paradigm
example

from 2.SG.PRES.IND to 1.SG.PRES.IND. The first step is to calculate a distance matrix based on the modified Levenshtein distance discussed before, this is shown in Table 7. For each form, I have highlighted the nearest neighbour.⁴² In this case, the nearest neighbour of *comes* is not *barres* but *tomas*. If we were to do the assignment solely based on this information, we would classify *comes* to the wrong class, namely $\langle X1, + \rangle o \rightleftharpoons \langle X1, + \rangle as$. However, we can do a filtering step, and remove proportions which are incompatible with the forms we are trying to classify. This step simply means narrowing the search space to those proportions which are real candidates for each lexeme in question. In this case, $\langle X1, + \rangle o \rightleftharpoons \langle X1, + \rangle as$ is incompatible with *comes* and thus we would correctly classify it as $\langle X1, + \rangle o \rightleftharpoons \langle X1, + \rangle es$, and the system would produce a perfect accuracy of 1.⁴³

	tocas	comes	barres	tomas
tocas	0	1.3	1.45	0.33
comes	1.03	0	0.95	0.7
barres	1.45	0.95	0	1.45
tomas	0.33	0.7	1.45	0

Table 7:
Distance Spanish
example

There is one thing worth mentioning. In the previous example, we used a right-hand-side weighted distance because we know Spanish is a suffixing language, and we know that the right-hand side of verbs is more important than the left-hand side for inflection class assignments. However, for most systems, we cannot know beforehand which of these three produces the best results for any given cell pair.

⁴²Since we only have 4 items in this example it makes no sense to use 5 nearest neighbours, but the same logic would apply if we consider more neighbours.

⁴³This example is too simple because the filtering is enough to get a perfect accuracy, but it helps illustrate the whole process.

For this reason, for each cell pair, we try all three (right-hand-side, left-hand-side, and average of both) and keep the best one (in terms of accuracy).⁴⁴

After having calculated the accuracy of all cell pairs in both directions, we take the average accuracy of the paradigm as the I-complexity of the paradigm.

Before discussing the results, I present a brief illustration on how edge-weighted Levenshtein distances compare to regular Levenshtein distances in a classification task, and we also compare these to a more general classifier, namely Boosting Trees with XGBoost (Chen and Guestrin 2016). For this comparison I picked 5 language datasets, with two cells for each dataset. The datasets in question are: Hungarian nouns, Latvian nouns, Yaitepec-Chatino verbs, Arabic verbs and Navajo verbs. I chose these datasets somewhat randomly, trying to maximise variety in terms of language families and paradigm structure. For each dataset, I first computed the proportions to go from one cell to the other as described in Section 3.2.1. I then performed the k -NN classification method described in this section using four different distance metrics: Levenshtein Distance (LD), right-hand-side edge-weighted LD (RHS), left-hand-side edge-weighted LD (LHS), and left-right-hand-side edge-weighted LD (LRHS). For all datasets, I computed the accuracy of predicting the inflection class of the pair from each cell.

Additionally, I trained a Boosting Tree classifier using XGBoost. Boosting Trees are a machine learning classification technique which consists of sequentially fitting small classification trees, and aggregating their predictions. Boosting Trees are similar in principle to Random Forest, with the difference that Random Forest fits multiple small classification trees randomly, while Boosting Trees work by sequentially fitting trees which target the errors in the previous tree. In practice, Boosting Trees have been successfully used in several classification tasks (Bonami and Pellegrini 2022; Guzmán Naranjo and

⁴⁴ A single language could use different similarities for different cell pairs. For example, if a cell pair analogy Cell 1 \rightleftharpoons Cell 2 is [<X1, + > . \rightleftharpoons . <X1, + >] then doing Cell 1 \rightarrow Cell 2 might work better with right hand side similarity (because it has a suffix) while doing Cell 2 \rightarrow Cell 1 might work better with left hand side similarity because it has a prefix.

Bonami 2021; Bonami *et al.* 2023), and they can perform extremely well. I used slightly different meta-parameters for each dataset, but the basic setup is that when predicting Cell 2 from Cell 1, I take the 5 to 7 final (or initial)⁴⁵ segments of Cell 1, and use them as predictors in the model. For each dataset, I optimised the meta-parameters with grid-search until the model achieved the best accuracy possible. In all cases, with k -NN and Boosting Trees, I performed 10-fold cross-validation.

Table 8 shows a comparison of these models. First, there are the results of k -NN using a simple Levenshtein distance and $k = 5$. Second, the results of k -NN with an edge-weighted Levenshtein distance and $k = 5$, the table shows results for the right-hand-side edge (RHS), left-hand-side edge (LHS), and left- and right-hand-side (LRHS) distances. Finally, it shows the results of a Boosting Tree algorithm trained on the N final (or initial) segments of the source inflected form, and the results of TiMBL fitted to the whole word.

The results show two key points. First, the accuracy of the edge-weighted Levenshtein distance models are systematically higher than the accuracies of the regular Levenshtein distance models, even if only by a small amount in some cases. The implication is that edge-weighting distances produce either equivalent, or better results in these five languages, and cell pairs which were chosen for their diverse structures. In some cases, like Hungarian and Latvian, the difference between regular LD and edge-weighted LD can be as dramatic as 11 percentage points. This performance difference is enough to justify preferring edge-weighted LD for our purpose. Second, and equally as important, the Boosting Tree classifier can outperform the distance-based k -NN classifiers most of the time,⁴⁶ and in some cases, by a very large margin, like in Yaitepec-Chatino or Navajo. This is perhaps not

⁴⁵I experimented with both sides, and chose the one which produced the best performance

⁴⁶The cases in which it does not, it reaches a very comparable accuracy. It is unclear why XGBoost sometimes struggles to outperform the k -NN classifiers, but here two factors are likely at play. First, machine learning techniques like XGBoost work better with larger datasets, and some of our datasets in this experiment are not very large (fewer than 1000 lexemes). Second, while I did my best to optimise the hyper-parameters of the models, it is possible that a different parametrization could produce in better results.

Table 8: Accuracy comparison classification methods

Language	POS	N. lexemes	N. classes	Predictor	Predicted	LD	RHS	LHS	LRHS	XGBT	TiMBL
Hungarian	N	11417	44	NOM.SG	ACC.PL	0.8	0.91	0.5	0.83	0.93	0.91
Hungarian	N	11417	44	ACC.PL	NOM.SG	0.97	0.98	0.9	0.94	0.98	0.95
Latvian	N	2515	43	NOM.SG	ACC.PL	0.85	0.91	0.5	0.85	0.93	0.92
Latvian	N	2515	43	ACC.PL	NOM.SG	0.85	0.93	0.6	0.8	0.95	0.95
YC	V	316	105	1CPL	3CPL	0.3	0.34	0.3	0.3	0.34	0.4
YC	V	316	105	3CPL	1CPL	0.4	0.44	0.4	0.4	0.69	0.49
Arabic	V	687	22	IMP.ACT.M/F.2.D	SBJV.ACT.F.3.D	0.97	0.92	0.96	0.97	0.96	0.94
Arabic	V	687	22	SBJV.ACT.F.3.D	IMP.ACT.M/F.2.D	0.96	0.92	0.96	0.95	0.96	0.94
Navajo	V	784	69	IPFV.1:IPA	IPFV.3r:IPA	0.89	0.64	0.93	0.9	0.91	0.85
Navajo	V	784	69	IPFV.3r:IPA	IPFV.1:IPA	0.78	0.63	0.79	0.73	0.95	0.77

surprising, as Boosting Trees can pick up much more complex patterns in the data. The implication is that the results I present in this paper are a complexity baseline, and that it is likely that with more time and computational resources one could fit models which result in higher accuracy and lower complexity than the ones I present here.

Taking stock, these results show that one word edge is clearly more important than the other edge for classification purposes, and that edge-weighted Levenshtein distances outperform regular Levenshtein distances; and also, that more sophisticated classification techniques should be able to produce better results.

RESULTS

4

This section presents the main results of this paper. It is divided into three subsections. First, I discuss the results for I-complexity, then the results of E-complexity, and finally I look at the relations between the two. One crucial fact to keep in mind is that these results should be interpreted as upper bounds on complexity. As I mentioned when discussing the classification method, it is likely that more advanced classifiers would produce lower I-complexity, but the same is true regarding E-complexity. A more sophisticated approach to inducing proportions could be able to reduce the fragmentation of many patterns by finding better and simpler abstractions.

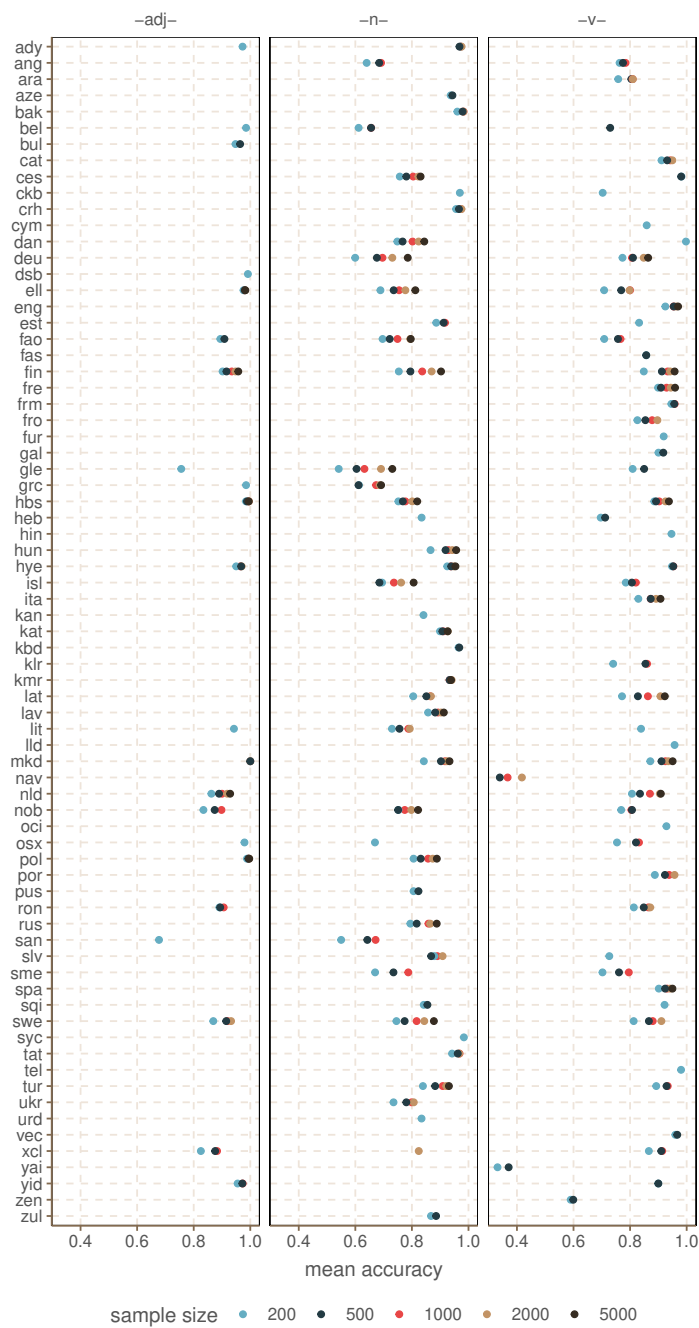
I-complexity

4.1

First, let us analyse the mean I-complexity across the whole paradigm of each language for each sample size.⁴⁷ These results are shown in Figure 1. The accuracy value for each dataset is the mean accuracy across all cell pair predictions. There are several important points worth mentioning here. First, most languages have accuracies above 0.7 for all sample sizes, for adjectives, nouns, and verbs, and

⁴⁷Recall that for all datasets, we created random subsamples of 200, 500, 1000, 2000, and 5000 lexemes. This allows us to better compare across all languages, for those datasets with very few lexemes.

Figure 1:
Mean accuracy
by language,
part of speech
and sample size



accuracies above 0.8 for sample sizes 2000 and 5000. Even with the very crude approach to classification taken in this paper, and even when only looking at 200 lexemes, we find that most systems have an accuracy above chance level. While these relatively high accuracies might seem unsurprising to linguists familiar with computational work on predicting class assignment of words, these results show that a very simple method for classification can achieve very high accuracies cross-linguistically, for many different types of inflectional systems, only based on phonological distances. It is important to note that these results should be understood as upper complexity limits. More sophisticated classification techniques like LSTMS are likely to be able to produce much better average accuracy scores. If the model reaches a mean accuracy of 96% for some language, this does not mean that the remaining 4% of cell pairs are unpredictable. Rather, it means that given the method and data we were only able to predict 96% of cell pairs. It is very likely that either a more sophisticated method like the ones mentioned in the background section, or more (e.g. simply more lexemes) and better data (e.g. semantic information), would allow us to reach a higher accuracy.

Another key point to remark on is that we are not choosing the best principal part for these results, but rather testing all possible cells and averaging across them. These results are averaged from the worst predictive cells and the best predictive cells. This observation connects to the second point, which is that inflection systems that are usually thought of as needing multiple principal parts, like Latin or Spanish verbs do not actually seem to need principal parts given that the mean accuracy is so high (>0.95 for 5000 lexemes). It is likely that some cells are very bad at predicting some other cells, but more often than not, knowing only one cell is enough to predict a good portion of the remaining cells as can be seen from the results.

If we were to pick the best predicting cell (akin to choosing the principal part), then the accuracy results can go up dramatically. For example, for Spanish verbs, the worst predictive cell in the 5000 sample size is the first singular present of the indicative with 0.86, while the best predictive cell is the infinitive with 0.98. Similarly, for Latin, the worst predictive cell in the 5000 sample size is FIN.IND.PRES.ACT.1.SING with a mean accuracy of 0.84, while the best predictive cell is FIN.IND.FUT.ACT.3.SING with a

mean predictive accuracy of 0.96. This fact further points toward the interpretation that I-complexity seems to be rather low cross-linguistically. Moreover, these results are an alternative way of measuring complexity similar to the principal parts approach, but without the challenges that are related to identifying principal parts already discussed.

A third point worth noting is that there is a very large amount of variation across languages. While some languages like Telugu (tel) have very low complexity in their verbal paradigm, others like Zenaga (zen) have a much larger complexity. There is also variation across domains for the same language. For example, Irish (gle) has a high complexity in the adjectival and nominal system, but lower complexity in the verbal system. These results do not show any clear tendency in terms of I-complexity across domains. While some languages are equally simple in all three domains (e.g. Armenian, hye), others are similarly complex across domains (e.g. Faroese, fao). The only clear trend appears to be that adjectives have lower complexity for this sample (although the sample has fewer adjective paradigms than verb or noun paradigms).

There are two clear exceptions to the high predictability result: Navajo (nav) and Yaitepec Chatino (yai). For languages like Navajo and Yaitepec Chatino, these results suggest that knowing just one cell is clearly not enough, and they raise the question of how many cells we need to know in these languages to be able to deduce the remaining cells. I discuss Navajo in some more detail next.

In the dataset, Navajo verbs can inflect for 7 persons: 1, 2, 3, 3o, 3a (fourth person), 3s (space), and 3i (indefinite);⁴⁸ 3 numbers: singular, dual and plural; and 5 TAM categories: future (FUT), imperfective (IPFV), iterative (ITER), optative (OPT) and perfective (PFV) (see Young 2000, for a more complete description of Navajo verbal inflection). Most verbs in our data have somewhere between 50 cells and 70 cells. Tables 9 and 10 show the inflection table for three verbs: *adika'* ('to play cards'), *náháshne* ('to hope around') and *yish'aah* ('to eat').

⁴⁸This is only a small fragment of Navajo verb conjugation, because the dataset only includes subject indices.

Table 9: Conjugation of *adika'* ('to play cards'), *náháshne* ('to hope around') and *yish'aah* ('to eat'), part 1

Tense	Person	Number	' <i>adiishk'</i> áq̣h	<i>náháshne'</i>	<i>yish'aah</i>
FUT	1	SG	?atite:ʃk'á:ʃ	nahote:ʃtʃ'á:h	te:ʃ?á:ʃ
FUT	1	DL	?atiti:k'á:ʃ	nahoti:tʃ'á:h	ti:ʃá:ʃ
FUT	1	PL	tati?ti:k'á:ʃ	ntahoti:tʃ'á:h	tati:ʃá:ʃ
FUT	2	SG	?atití:k'á:ʃ	nahotí:tʃ'á:h	tí:ʃá:ʃ
FUT	2	DL	?atito:hk'á:ʃ	nahoto:hʃ'á:h	to:hʃá:ʃ
FUT	2	PL	tati?to:hk'á:ʃ	ntahoto:hʃ'á:h	tato:hʃá:ʃ
FUT	3	SG	?atito:k'á:ʃ	nahoto:tʃ'á:h	to:ʃá:ʃ
FUT	3	PL	tati?to:k'á:ʃ	ntahoto:tʃ'á:h	tato:ʃá:ʃ
FUT	3a	SG	?aʒtito:k'á:ʃ	nahozto:tʃ'á:h	tʃito:ʃá:ʃ
FUT	3a	PL	tatiʒ?to:k'á:ʃ	ntahozto:tʃ'á:h	taʒto:ʃá:ʃ
IPFV	1	SG	?ati:ʃk'á:h	nahaʃtʃ'á:h	jiʃ?a:h
IPFV	1	DL	?ati:k'á:h	nahwi:tʃ'á:h	ji:ʃa:h
IPFV	1	PL	ta?ti:k'á:h	ntahwi:tʃ'á:h	tei:ʃa:h
IPFV	2	SG	?ati:k'á:h	nahótʃ'á:h	ni?a:h
IPFV	2	DL	?ato:hk'á:h	nahohʃ'á:h	woh?a:h
IPFV	2	PL	ta?to:hk'á:h	ntahohʃ'á:h	ta:h?a:h
IPFV	3	SG	?ati:k'á:h	nahaʃ'á:h	ji?a:h
IPFV	3	PL	ta?ti:k'á:h	natahaʃ'á:h	ta:ʃa:h
IPFV	3a	SG	?aʒti:k'á:h	nahotʃitʃ'á:h	tʃi?a:h
IPFV	3a	PL	taʒ?ti:k'á:h	ntahotʃitʃ'á:h	taʃi?a:h
ITER	1	SG	ń?ti:ʃk'á:h	nináháʃtʃ'á:h	náʃ?á:ʃ
ITER	1	DL	ń?ti:k'á:h	nináhwi:tʃ'á:h	néi:ʃá:ʃ
ITER	1	PL	ńta?ti:k'á:h	ninátahwi:tʃ'á:h	ńtei:ʃá:ʃ
ITER	2	SG	ń?ti:k'á:h	nináhótʃ'á:h	nání?á:ʃ
ITER	2	DL	ń?to:hk'á:h	nináhóhʃ'á:h	náh?á:ʃ
ITER	2	PL	ńta?to:hk'á:h	ninátahohtʃ'á:h	ńta:hʃá:ʃ
ITER	3	SG	ń?ti:k'á:h	nináháʃtʃ'á:h	náʃá:ʃ

Table 10:
Conjugation
of *adika'*
(‘to play cards’),
náháshne (‘to
hope around’)
and *yish’aah*
(‘to eat’), part 2

Tense	Person	Number	<i>’adiishk’áqah</i>	<i>náháshne’</i>	<i>yish’aah</i>
ITER	3	PL	ńtaʔti:k’á:h	ninátahatʃ’áh	ńta:ʔá:h
ITER	3a	SG	ńíʒʔti:k’á:h	nináhohʃitʃ’áh	ńtʃíʔá:h
ITER	3a	PL	ńtaʒʔti:k’á:h	ninátahotʃitʃ’áh	ńtatʃiʔá:h
OPT	1	SG	ʔato:ʃk’á:ʃ	nahóʃʃ’á:h	wóʃʔá:ʃ
OPT	1	DL	ʔato:k’á:ʃ	nahó:ʃ’á:h	wo:ʔá:ʃ
OPT	1	PL	taʔto:k’á:ʃ	ntahó:ʃ’á:h	tao:ʔá:ʃ
OPT	2	SG	ʔatoók’á:ʃ	nahó:ʃ’á:h	wó:ʔá:ʃ
OPT	2	DL	ʔato:hk’á:ʃ	nahó:hʃ’á:h	wo:hʔá:ʃ
OPT	2	PL	taʔto:hk’á:ʃ	ntahó:hʃ’á:h	tao:hʔá:ʃ
OPT	3	SG	ʔato:k’á:ʃ	nahóʃ’á:h	wóʃʔá:ʃ
OPT	3	PL	taʔto:k’á:ʃ	ntahóʃ’á:h	taoʔá:ʃ
OPT	3a	SG	ʔaʒto:k’á:ʃ	nahóʃóʃ’á:h	ʃóʔá:ʃ
OPT	3a	PL	taʒʔto:k’á:ʃ	ntahotʃóʃ’á:h	taʃóʔá:ʃ
PFV	1	SG	ʔati:ʃk’á:ʔ	nahóʃéʃ’á:ʔ	jíʔá
PFV	1	DL	ʔati:ʃk’á:ʔ	nahóʃi:ʃ’á:ʔ	ji:ʔá
PFV	1	PL	taʔti:ʃk’á:ʔ	ntahóʃi:ʃ’á:ʔ	tei:ʔá
PFV	2	SG	ʔatiniʃk’á:ʔ	nahosíníʃ’á:ʔ	jíníʔá
PFV	2	DL	ʔato:hʃk’á:ʔ	nahóʃo:ʃ’á:ʔ	wo:ʔá
PFV	2	PL	taʔto:hʃk’á:ʔ	ntahóʃo:ʃ’á:ʔ	tao:ʔá
PFV	3	SG	ʔati:ʃk’á:ʔ	nahazʃ’á:ʔ	jíʔá
PFV	3	PL	taʔti:ʃk’á:ʔ	ntahazʃ’á:ʔ	tá:ʔá
PFV	3a	SG	ʔaʒti:ʃk’á:ʔ	nahotʃiʒʃ’á:ʔ	ʃí:ʔá
PFV	3a	PL	taʒʔti:ʃk’á:ʔ	ntahotʃiʒʃ’á:ʔ	taʃí:ʔá

The main difficulty in Navajo seems to come from predicting across TAM categories. Measuring the predictability within TAM blocks (PFV cells only predicted from other PFV cells, etc.), the mean accuracy of the system is 0.81, which is clearly much better than the 0.42 of the mean accuracy of the whole system. Looking at the best predictors for each TAM block we get the results in Table 11. This means that in Navajo, it is relatively easy to predict all cells of a verb as long as you know one form for each of these blocks. Even taking the worst predictors by TAM in Navajo, as shown in Table 12, the system still has very high inter-predictability.

TAM	predictor	accuracy
FUT	FUT.3.SG	0.947
IPFV	IPFV.3.SG	0.935
ITER	ITER.3.SG	0.952
OPT	OPT.3.SG	0.962
PFV	PFV.3.SG	0.863

Table 11:
Best predictors by TAM category
for Navajo

TAM	predictor	accuracy
FUT	FUT.3I.SG	0.805
IPFV	IPFV.1.PL	0.707
ITER	ITER.1.PL	0.740
OPT	OPT.1.PL	0.731
PFV	PFV.2.SG	0.645

Table 12:
Worst predictors by TAM category
for Navajo

This does not quite mean that Navajo necessarily requires five principal parts. Table 13 shows that for FUT, IPFV, and OPT, the model gets a relatively high accuracy from at least one cell from a different TAM block. These results come from choosing the best predictor found in a different TAM block. For FUT, the accuracy is lower (0.81), and for PFV the accuracy is very low (0.403). This shows that the main difficulty comes from predicting PFV from non-PFV cells.

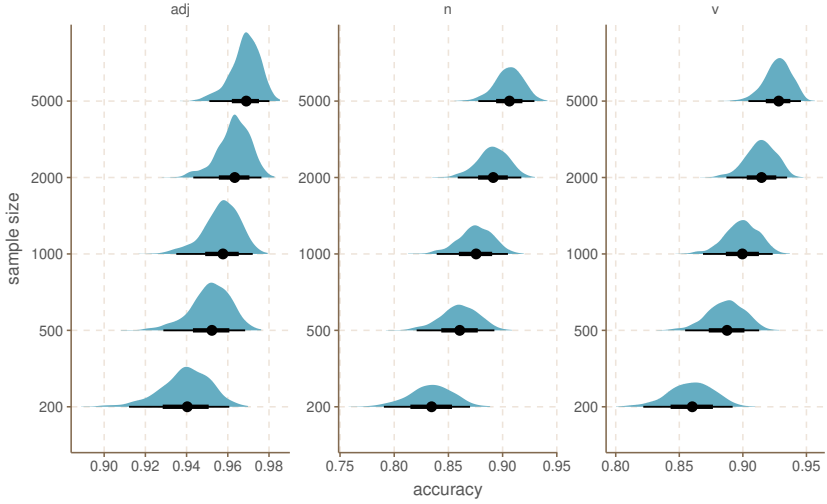
What the Navajo example shows is that even apparently very complex systems like Navajo have only limited I-complexity in the sense that this complexity is mostly restricted to predicting across certain TAM features, and it is not a general property of the whole system.

A different, but equally important question is whether the number of lexemes in a corpus impacts our estimates of I-complexity. I built a

predictor	predicted	accuracy
ITER.3S.SG	FUT.3I.SG	0.816
OPT.3I.SG	IPFV.3S.SG	0.846
FUT.3S.SG	ITER.1.DL	0.710
IPFV.3S.SG	OPT.3O.PL	0.803
OPT.3O.PL	PFV.3.PL	0.403

Table 13:
Best predictors across TAM categories
for Navajo

Figure 2:
Mean accuracy
vs sample size



Bayesian⁴⁹ zero-one inflated Beta regression model where to predict the mean accuracy from the sample size and the part of speech (verb, noun, or adjective) and controls for language by part of speech.⁵⁰ Figure 2 shows the marginal effects of sample size on the mean accuracy. Overall, there is an effect of sample size on mean accuracy, but this effect is relatively small, especially for adjectives. For verbs and nouns, the model does not show any noticeable difference between 2000 and 5000 lexemes, but the difference between 200 and 5000 is more clear. While having larger sample sizes can lead to higher accuracy estimates, it is not clear that relatively small number of lexemes produce bad estimates. Moreover, we can be confident that higher sample sizes lead to higher mean accuracy, meaning that estimates on small sample sizes work well as a lower bound. The consequence is that we can study I-complexity for languages using this method even if we only have access to relatively small datasets. This is a key result. This method allows us to study the I-complexity in languages with

⁴⁹I used Stan (Carpenter *et al.* 2017) with brms Bürkner (2017) for all models in this paper.

⁵⁰The formula in question in brms is `mean-accuracy ~ mo(sample_size) * pos + (1 + mo(sample_size) | language/pos)`, where `mo` is a function to declare monotonic effects.

considerably smaller resources than those needed when using LSTMs (Cotterell *et al.* 2019).

E-complexity

4.2

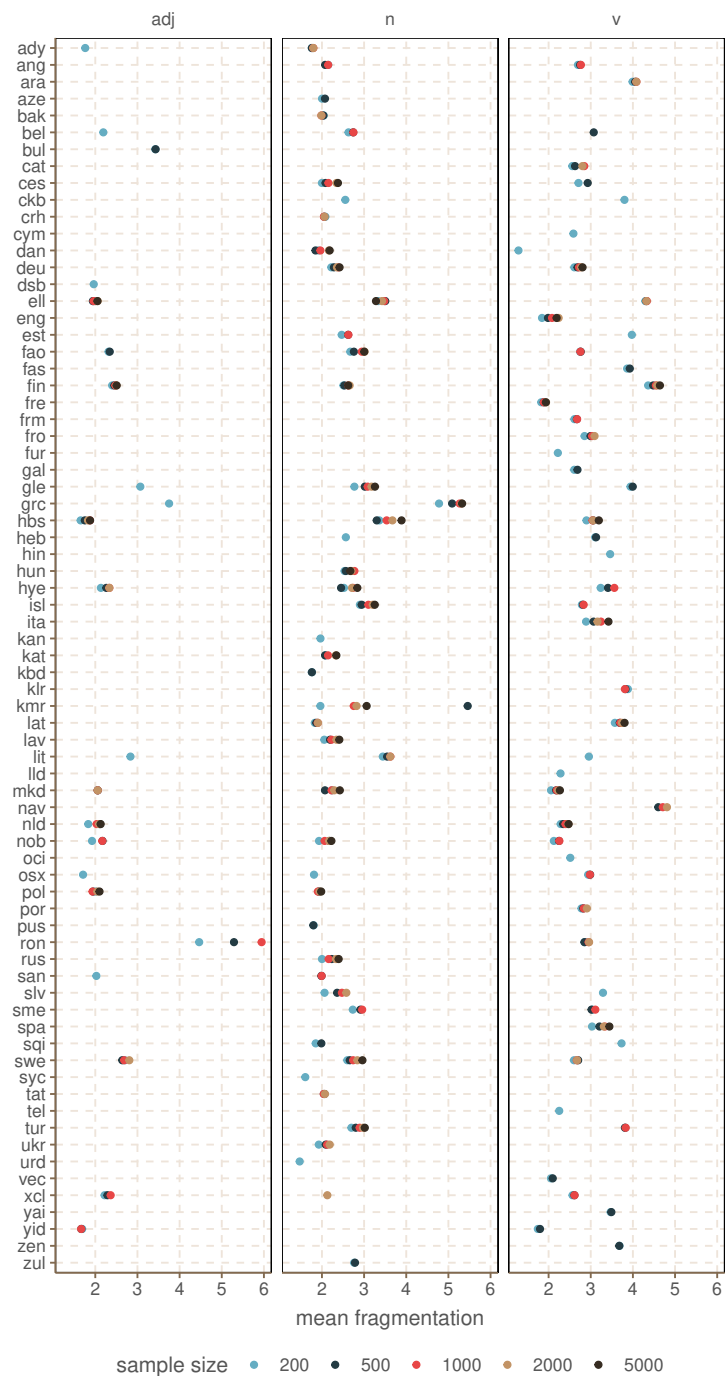
As introduced in Section 3.2.1, we measure E-complexity in terms of the fragmentation of the proportions between cells. Figure 3 shows the mean fragmentation by language and sample size. For the most part, fragmentation stays relatively stable across sample sizes except for Northern Kurdish (kmr).⁵¹

At the same time, the mean fragmentation for most languages is higher than 2. This result shows that inflectional pairs of the form *stem + ending 1:stem + ending 2* are not the most common pattern in our sample, and it shows that it is much more common to have at least two breaks in inflectional pairs. This is even true in European languages which are usually analysed in segmentation-based approaches as being composed of a (mostly invariant) stem and an ending. This does not seem to be the most common situation on average. While these fragmentation values are dependent on the chosen formalization of proportions, they do suggest that for studies of inflectional complexity methods which focus on suffixes and prefixes, and ignore alternations within inflected words, could underestimate E-complexity.

Another important implication of these results is that approaches which follow segmentation based on linguistic traditions, and which are not designed to be language-independent, are likely to overestimate the complexity of some languages, and to underestimate the complexity of other languages. To illustrate this point, we can look at the fragmentation of Arabic (ara), Spanish (spa), and English (eng) verbs. The mean fragmentation of English verbs is of approximately 2.1, for Spanish verbs in the 5000 verb sample is of approximately 3.5, and the mean fragmentation of Arabic verbs is of about 4. However, if one simply follows traditional descriptions of these three languages, Arabic is often characterised as having triconsonantal stems with different affixing schemas, while Spanish and English are characterised as being a stem + ending type of languages. Our results show that the

⁵¹The large difference in fragmentation of Kurdish between the smaller and larger datasets is due to a subset of lexemes with additional periphrastic cells.

Figure 3:
Mean
fragmentation
by language
and sample size



difference (at least in terms of E-complexity) between Arabic on the one hand, and Spanish and English on the other, is not a categorical one, but rather a gradient one. While Arabic is in fact more complex than English and Spanish, Spanish is much closer to Arabic than it is to English.

Unlike I-complexity, we do find large variation in the E-complexity of different languages, anywhere between 2 and 6 mean fragmentation. This, despite the fact that our method treats all types of *stem changes* in the same way. The result shows that some languages make use of substantially more discontinuous markers (i.e. markers which happen at separate positions) than others.

As with accuracy, we are interested in exploring how sample sizes affect our estimates of fragmentation. I fitted a log-normal model⁵² with the same predictors as for accuracy.⁵³ The result is similar to the others for accuracy, but the effect goes in the opposite direction in terms of complexity, that is, the larger the sample size, the higher the E-complexity. This can be seen in Figure 4. Smaller sample sizes under-

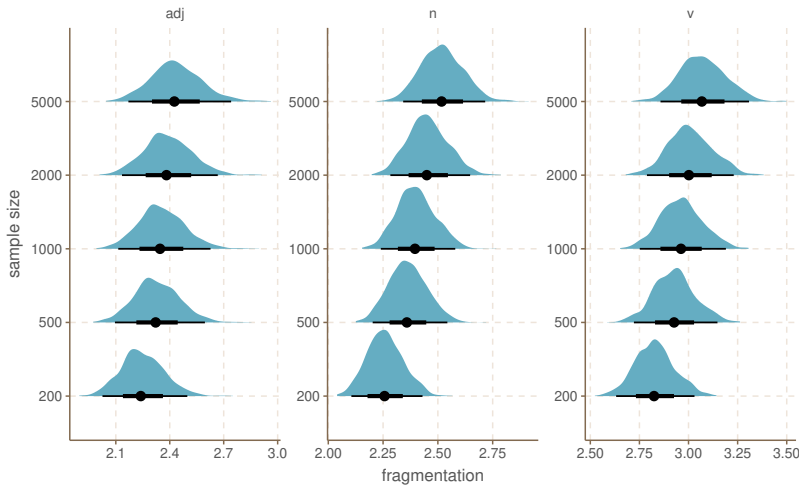


Figure 4:
Mean
fragmentation
vs sample size

estimate the E-complexity of the system, but the effect is very small.

⁵²I used a log-normal likelihood because our mean fragmentation values can only be positive.

⁵³As before: $\text{accuracy} \sim \text{mo}(\text{sample_size}) * \text{pos} + (1 + \text{mo}(\text{sample_size}) \mid \text{language}/\text{pos})$.

Even at only 200 lexemes, the estimates are very close to the estimates with 5000 lexemes. The likely explanation for this effect is that larger samples contain more unique inflection patterns, or suppletive forms which increase the mean fragmentation of the system.

4.3

E- and I-complexity trade-offs

A question that has been asked multiple times in typology is the relation between the complexity of different parts of a grammatical system. With respect to morphology in particular, Cotterell *et al.* (2019) propose a negative correlation between E- and I-complexity. Namely, the authors find that as I-complexity increases, E-complexity decreases, and the other way around. Cotterell *et al.* (2019) use a LSTM approach to estimate the I-complexity of 36 languages, and paradigm size as a measure of E-complexity.⁵⁴ The implication is then that there is effectively a trade-off in terms of complexity, and thus, arguably, a sort of upper level of complexity for any inflectional system.

First, we want to compare the I-complexity results against E-complexity measured in terms of paradigm size. Figure 5 shows the mean accuracy by language and part of speech vs the number of cells in the relevant paradigm.⁵⁵ Unlike in the case of results reported by Cotterell *et al.* (2019), there does not appear to be any type of correlation between I-complexity and the number of cells. There are two possible reasons for this discrepancy in results. One possibility is that our approach to measuring I-complexity just does not show the type of correlation that Cotterell *et al.* (2019) found. While this is possible, it is not possible to test this explanation without direct access to the original dataset used in that paper.⁵⁶ The alternative is that there is bias in the dataset used by Cotterell *et al.* (2019), and that a larger dataset removes any sort of bias of their smaller dataset.

⁵⁴However, Cotterell *et al.* (2019) only count the number of different cell realisations, rather than total number of cells listed.

⁵⁵Since some paradigms have a small amount of variation in the number of cells a lexeme allows depending on the type of lexeme, I take the maximum possible number of cells.

⁵⁶Cotterell *et al.* (2019) also use UniMorph data, but it is not completely clear which version was used, because these datasets have seen changes since the original study was published.

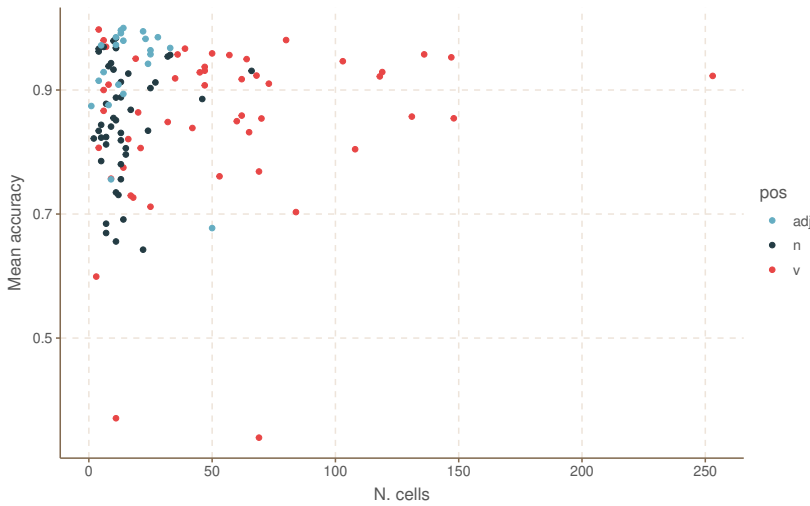
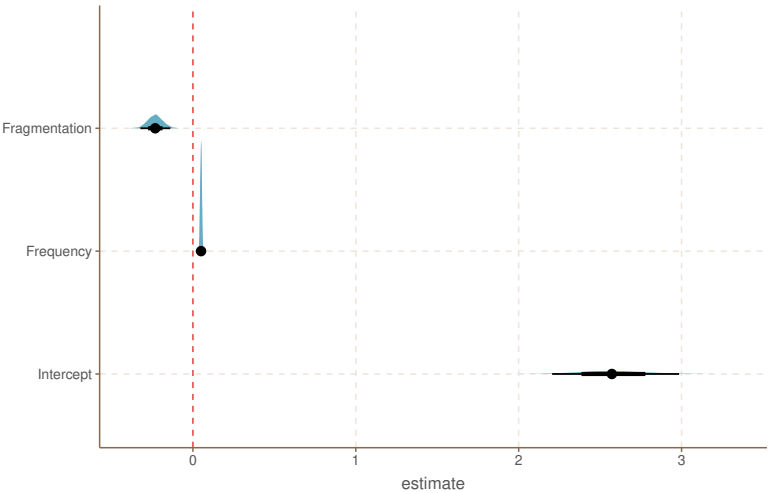


Figure 5:
Mean accuracy
vs number
of cells

However, this type of approach assumes that the relation between E- and I-complexity happens at the level of the whole inflectional system. There is no apriori reason why this should be the case, however. It is more likely that the correlation, if any, should happen at the individual pattern level. For example, it is possible that relatively simple suffixing proportions with low fragmentation like $X \rightleftharpoons Xa$ will also be easier to predict than more complex proportions with higher fragmentation like $uXiY \rightleftharpoons XaYo$. To test this hypothesis, I fitted a binomial model predicting the accuracy of each pattern from its fragmentation and controlled for language. Because there is so much data, and so many proportions, I had to downsample the dataset.⁵⁷ First, I restrict the model to results from the datasets with 1000 lexemes. Additionally, since verbs can have many cells (sometimes in the hundreds), I took a random sample of 500 proportions per language for the verb dataset which left us with around 10,000 proportions instead of 100,000 (about 6000 for nouns and 1800 for adjectives). This leaves us with a smaller dataset, which should still contain enough information to allow us to estimate any effects in the data.

⁵⁷ The issue arises because fitting the group-level effects with a correlation structure is very slow and difficult (i.e. the funnel geometry of the space leads to divergences in the sampling).

Figure 6:
Coefficients
of the model
comparing
pattern accuracy
vs fragmentation



I fitted a binomial model predicting pattern accuracy from its fragmentation. The question boils down to: is there a relation between the I-complexity of a pattern and its E-complexity? I also controlled for the type frequency of the pattern in the cell,⁵⁸ as well as language and part of speech.⁵⁹ The main coefficients for the model are shown in Figure 6. The results show a negative effect of the proportion's fragmentation on the model's accuracy predicting it, and, as expected, a clear positive effect of frequency on accuracy, meaning that more frequent proportions are easier to predict than less frequent ones. Since accuracy is the opposite of complexity, it means that a higher fragmentation in a pattern generally leads to higher complexity. This result is effectively the opposite of a complexity trade-off. More complex proportions in terms of E-complexity also tend to be harder to predict, while simpler proportions tend to be easier to predict.

Understanding why this effect happens in our data is not completely straightforward, but there are some potential explanations.

⁵⁸ A very frequent pattern, i.e. a pattern that applies to many lexemes in a cell, could be easier to predict than a rarer one. The frequency of a pattern by cell could be correlated with its complexity.

⁵⁹ The brms model was the following: `correct | trials(total) ~ 1 + fragmentation + log(total) + (1 + fragmentation | language/pos)`

If more complex proportions are harder to predict, a reasonable hypothesis is that the number of infixes in the proportions might be driving this effect. To test this, we first look at the mean number of infixes in low and high accuracy proportions. I looked at the results from the datasets with 1000 lexemes. From these, I then extracted the 100 proportions with highest accuracy, and the 100 proportions with lowest accuracy for each language. The results shown in Figure 7. The pattern is clear: the most accurate proportions systematically have the same or fewer number of infixes than the least accurate proportions.

Next, we can approach the question from the opposite direction and only look at the best performing proportions. For this, I further restricted the sample to proportions with a frequency of between 2 and 100 (to control for effects of very high frequent proportions). I also abstracted away all concrete material and matching potential to get basic skeletal patterns: [$\langle X \rangle . \Leftarrow \langle X \rangle .$]. Then, I extracted the 10 most frequent proportions among those with an accuracy of 1, those with an accuracy higher than 0.95, and those with an accuracy higher than 0.9, and then compare their relative frequency in those subsamples to their relative frequency in the whole dataset. I did this experiment aggregating across all languages. The results of this comparison are shown in Table 14. By comparing the values in columns ‘acc=1’, ‘acc>0.95’, ‘acc>0.9’ to the values in the baseline column ‘total sample’, one can see the extent to which a pattern is over-represented among the most-accurate proportions, relative to its overall frequency.

Out of the 10 most frequent proportions on the three subsamples, only [$\langle X \rangle . \Leftarrow \langle X \rangle .$] shows any clear difference in relative frequency between the subsample and the whole sample, but this difference is considerable. For the subsample on proportions with accuracy of 1, this takes up 0.1 additional total frequency than in the whole sample. What this mean is that the proportion [$\langle X \rangle . \Leftarrow \langle X \rangle .$] is very common among easy to predict proportions, while we observe more proportions with more infixes among the harder to predict proportions. This helps create the observed correlation between E- and I-complexity.

Figure 7:
Mean infix
accuracy

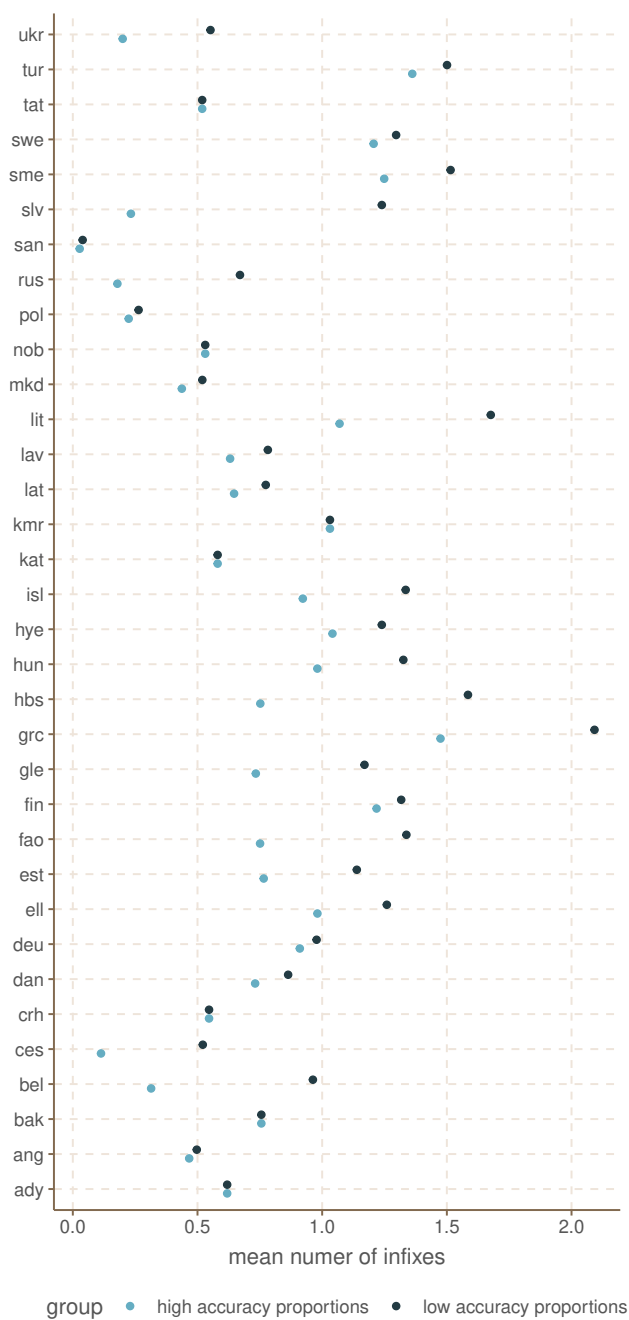


Table 14: Relative frequency of most accurate proportions

Proportion	acc = 1	acc > 0.95	acc > 0.9	Total sample
$\langle X \rangle . \rightleftharpoons \langle X \rangle .$	0.49	0.44	0.43	0.39
$\langle X \rangle . \langle X \rangle . \rightleftharpoons \langle X \rangle . \langle X \rangle .$	0.09	0.10	0.10	0.09
$\langle X \rangle . \langle X \rangle \rightleftharpoons \langle X \rangle . \langle X \rangle$	0.07	0.07	0.07	0.07
$\langle X \rangle . \langle X \rangle \rightleftharpoons \langle X \rangle . \langle X \rangle .$	0.06	0.06	0.06	0.05
$\langle X \rangle . \langle X \rangle . \rightleftharpoons \langle X \rangle . \langle X \rangle$	0.04	0.05	0.05	0.05
$\langle X \rangle . \rightleftharpoons \langle X \rangle$	0.03	0.03	0.03	0.03
$\langle X \rangle \rightleftharpoons \langle X \rangle .$	0.02	0.02	0.03	0.03
$\langle X \rangle . \langle X \rangle . \langle X \rangle \rightleftharpoons \langle X \rangle . \langle X \rangle . \langle X \rangle$	0.02	0.02	0.02	0.02
$\langle X \rangle . \langle X \rangle . \rightleftharpoons \langle X \rangle \langle X \rangle .$	0.01	0.01	0.02	0.02
$\langle X \rangle \langle X \rangle . \rightleftharpoons \langle X \rangle . \langle X \rangle .$	0.01	0.01	0.01	0.02

CONCLUSION

5

In this paper, I have presented an approach to the typology of paradigm complexity in the spirit of Word and Paradigm morphology. I argue that a W&P approach is advantageous for doing cross-linguistic work in inflectional morphology for multiple reasons. First, it gets around the segmentation problem, and second, it allows for relatively simple formalisation in the form of proportional analogies that can be used for efficient automatic induction. I have presented a concrete formalisation of proportional analogies, using named variables with matching potential, restricting morphological patterns to be defined from the word boundary. With this formalisation, I have shown that it is possible to measure both E- and I-complexity in many typologically diverse morphological systems.

The results confirm previous results in the literature (Ackerman and Malouf 2013). The I-complexity of most morphological systems examined were relatively low, and increasing sample sizes leads to a reduction in system complexity. In contrast, E-complexity is less consistent across languages and parts of speech. The results also show that there is a clear correlation between I- and E-complexity of individual patterns: patterns with higher E-complexity lead to higher I-complexity. At the same time, there does not seem to be a clear cor-

relation between I-complexity and paradigm size as has been reported in the literature. The lack of a trade-off between different levels of morphological complexity also point towards the conclusion that, among morphologically complex languages, some are decidedly more complex than others.

There are also some wider implications for the study of morphological typology in general. Using automatic induction has the advantage of being neutral to linguistic tradition, and it allows for systematic and comparable analysis for different languages. The fact that some languages have traditionally been described as using root-and-pattern morphology, or suffixes plus phonological rules for stem alternation, does not play a role in this approach since we analyse everything from a purely surface-based perspective. This is important because it is a fundamental requirement to be able to carry out large scale quantitative studies of morphological systems.

From a methodological perspective, this paper offers two contributions. First, I have shown that computational work in inflectional morphology is feasible with a relatively small number of lexemes. While this is not a completely new insight, it is important to emphasise this point. The fact that data is somewhat limited for many languages does not mean that we need to exclude them in computational approaches to morphology, it just means that we need to use tools capable of coping with small datasets. Second, I provided a new implementation of proportional analogies based on a new formalism. I have shown one potential application of this method to the estimation of inflectional complexity, but other applications are possible, and there is potential for further research on automated morphological analysis. At the same time, while this new formalism can capture a relatively wide range of phenomena, there are still some gaps which we aim to cover in future work, like inducing different types of reduplication and implementing feature structures.

ACKNOWLEDGMENTS

This research was partly funded by the Emmy Noether project Bayesian modelling of spatial typology (grant no. GU 2369/1-1, project number 504155622). I am grateful to Olivier Bonami for his many comments and suggestions.

REFERENCES

- Farrell ACKERMAN and Robert MALOUF (2013), Morphological organization: the low conditional entropy conjecture, *Language*, 89(3):429–464, doi:10.1353/lan.2013.0054.
- Adam ALBRIGHT, Argelia ANDRADE, and Bruce HAYES (2001), Segmental environments of Spanish diphthongization, *UCLA Working Papers in Linguistics*, 7(5):117–151.
- Adam ALBRIGHT and Bruce HAYES (1999), An automated learner for phonology and morphology, <https://pdfs.semanticscholar.org/8d74/847ecd575887fcfe42ea022c2d82750fe7d9.pdf>, unpublished manuscript.
- Peter ARKADIEV and Francesco GARDANI (2020), The complexities of morphology, in Peter ARKADIEV and Francesco GARDANI, editors, *The complexities of morphology*, pp. 1–19, Oxford University Press.
- Sabine ARNDT-LAPPE (2011), Towards an exemplar-based model of stress in English noun–noun compounds, *Journal of Linguistics*, 47(3):549–585.
- Sabine ARNDT-LAPPE (2014), Analogy in suffix rivalry: the case of English *-ity* and *-ness*, *English Language and Linguistics*, 18(3):497–548.
- R. Harald BAAYEN, Yu-Ying CHUANG, and Maria HEITMEIER (2019a), WpmWithLdl: implementation of word and paradigm morphology with linear discriminative learning R package version 2.
- R. Harald BAAYEN, Yu-Ying CHUANG, Elnaz SHAF AEI-BAJESTAN, and James P. BLEVINS (2019b), The discriminative lexicon: a unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning, *Complexity*, 2019:1–39.
- R. Harald BAAYEN, Richard PIEPENBROCK, and Leon GULIKERS (1996), The CELEX lexical database (cd-rom).

- Matthew BAERMAN, Dunstan BROWN, and Greville G. CORBETT, editors (2015), *Understanding and measuring morphological complexity*, Oxford University Press.
- Matthew BAERMAN, Dunstan BROWN, and Greville G. CORBETT (2017), *Morphological complexity*, Cambridge University Press.
- Sacha BENIAMINE (2017), Un algorithme universel pour l'abstraction automatique d'alternances morphophonologiques, in *24e conférence sur le traitement automatique des langues naturelles (TALN)*, volume 2.
- Sacha BENIAMINE (2018), *Classifications flexionnelles: étude quantitative des structures de paradigmes*, Ph.D. thesis, Université Paris Diderot.
- Sacha BENIAMINE (Forthcoming), One lexeme, many classes: inflection class systems as lattices, in Berthold CRYSMANN and Manfred SAILER, editors, *One-to-many relations in morphology, syntax and semantics*, Language Science Press.
- Sacha BENIAMINE, Olivier BONAMI, and Ana R. LUÍS (2021), The fine implicative structure of European Portuguese conjugation, *Isogloss. Open Journal of Romance Linguistics*, 7:1–35, ISSN 2385-4138, doi:10.5565/rev/isogloss.109, <https://revistes.uab.cat/isogloss/article/view/v7-beniamine-bonami-luis>.
- Sacha BENIAMINE and Matías GUZMÁN NARANJO (2021), Multiple alignments of inflectional paradigms, in *Proceedings of the Society for Computation in Linguistics (SCiL)*, volume 4, pp. 216–227, doi:10.7275/ymc0-p491.
- Charles Edwin BENNETT (1918), *New Latin grammar*, Allyn and Bacon.
- Christian BENTZ and Dimitrios ALIKANIOTIS (2016), The word entropy of natural languages, unpublished arXiv manuscript.
- Christian BENTZ, Ximena GUTIERREZ-VASQUES, Olga SOZINOVA, and Tanja SAMARDŽIĆ (2022), Complexity trade-offs and equi-complexity in natural languages: a meta-analysis, *Linguistics Vanguard*, doi:10.1515/lingvan-2021-0054.
- Christian BENTZ, Tatyana RUZSICS, Alexander KOPLÉNIG, and Tanja SAMARDŽIĆ (2016), A comparison between morphological complexity measures: typological data vs. language corpora, in *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)*, pp. 142–153.
- Christian BENTZ and Bodo WINTER (2013), Languages with more second language learners tend to lose nominal case, *Language Dynamics and Change*, 3(1):1–27, doi:10.1163/22105832-13030105.
- Balthasar BICKEL, Goma BANJADE, Martin GAENZLE, Elena LIEVEN, Netra Prasad PAUDYAL, Ichchha Purna RAI, Manoj RAI, Novel Kishore RAI, and Sabine STOLL (2007), Free prefix ordering in Chintang, *Language*, pp. 43–73.

- Balthasar BICKEL and Johanna NICHOLS (2007), Inflectional morphology, in Timothy SHOPEN, editor, *Language typology and syntactic description*, volume 3, pp. 169–240, Cambridge University Press, 2 edition.
- Balthasar BICKEL and Johanna NICHOLS (2013), Inflectional synthesis of the verb, in Matthew S. DRYER and Martin HASPELMATH, editors, *The world atlas of language structures online*, Max Planck Digital Library.
- James P. BLEVINS (2006), Word-based morphology, *Journal of Linguistics*, 42(3):531–573.
- James P. BLEVINS (2013), The information-theoretic turn, *Psihologija*, 46(3):355–375, ISSN 00485705, doi:10.2298/PSI1304355B.
- James P. BLEVINS (2016), *Word and paradigm morphology*, Oxford University Press.
- Olivier BONAMI and Sacha BENIAMINE (2016), Joint predictiveness in inflectional paradigms, *Word Structure*, 9(2):156–182.
- Olivier BONAMI and Sacha BENIAMINE (2021), Leaving the stem by itself, in Marcia HAAG, Sedigheh MORADI, Andrija PETROVIC, and Janie REES-MILLER, editors, *All things morphology*, pp. 82–98, John Benjamins.
- Olivier BONAMI, Gauthier CARON, and Clément PLANCQ (2014), Construction d'un lexique flexionnel phonétisé libre du Français, in *SHS web of conferences*, volume 8, pp. 2583–2596, EDP Sciences, doi:10.1051/shsconf/20140801223.
- Olivier BONAMI and Berthold CRYSMANN (2013), Morphotactics in an information-based model of realisational morphology, in Stefan MÜLLER, editor, *Proceedings of the 20th international conference on Head-Driven Phrase Structure Grammar*, Freie Universität Berlin, pp. 27–47.
- Olivier BONAMI, Lukáš KYJÁNEK, and Marine WAUQUIER (2023), Assessing the featural organisation of paradigms with distributional methods, *Proceedings of the Society for Computation in Linguistics*, 6(1):310–320.
- Olivier BONAMI and Matteo PELLEGRINI (2022), Derivation predicting inflection: a quantitative study of the relation between derivational history and inflectional behavior in Latin, *Studies in Language*, 46(4):753–792, doi:10.1075/sl.21002.bon.
- Joan L. BYBEE and Dan I. SLOBIN (1982), Rules and schemas in the development and use of the English past tense, *Language*, 58(2):265–289.
- Paul-Christian BÜRKNER (2017), Brms: an R package for bayesian multilevel models using stan, *Journal of Statistical Software*, 80(1):1–28, doi:10.18637/jss.v080.i01.
- Franco Alberto CARDILLO, Marcello FERRO, Claudia MARZI, and Vito PIRRELLI (2018), Deep learning of inflection and the cell-filling problem, *IJCoL. Italian Journal of Computational Linguistics*, 4(4-1):57–75.

- Bob CARPENTER, Andrew GELMAN, Matthew HOFFMAN, Daniel LEE, Ben GOODRICH, Michael BETANCOURT, Marcus BRUBAKER, Jiqiang GUO, Peter LI, and Allen RIDDELL (2017), Stan: a probabilistic programming language, *Journal of Statistical Software, Articles*, 76(1):1–32, ISSN 1548-7660, doi:10.18637/jss.v076.i01.
- Andrew CARSTAIRS (1983), Paradigm economy, *Journal of Linguistics*, 19(1):115–128.
- Andrew CARSTAIRS (1990), Phonologically conditioned suppletion, in Wolfgang U. DRESSLER, Hans C. LUSCHÜTZKY, Oskar E. PFEIFFER, and John R. RENNISON, editors, *Contemporary morphology*, number 49 in Trends in Linguistics, pp. 17–23, De Gruyter, Berlin.
- Andrew CARSTAIRS (1998), Some implications of phonologically conditioned suppletion, in Geert E. BOOIJ and Jaap VAN MARLE, editors, *Yearbook of morphology 1998*, pp. 67–94, Springer.
- Andrew CARSTAIRS-MCCARTHY (1994), Inflection classes, gender, and the principle of contrast, *Language*, pp. 737–788.
- Tianqi CHEN and Carlos GUESTRIN (2016), Xgboost: a scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794.
- Ryan COTTERELL, Christo KIROV, Mans HULDEN, and Jason EISNER (2019), On the complexity and typology of inflectional morphological systems, *Transactions of the Association for Computational Linguistics*, 7:327–342, doi:10.1162/tacl_a_00271.
- Sara COURT, Micha ELSNER, and Andrea D. SIMS (2022), Quantifying factors shaping analogical restructuring of the Maltese nominal system, Talk at the International Morphology Meeting, Budapest.
- Michael A. COVINGTON and Joe D. MCFALL (2008), The moving-average type-token ratio, in *Linguistics Society of America*.
- Michael A. COVINGTON and Joe D. MCFALL (2010), Cutting the Gordian knot: the moving-average type–token ratio (MATTR), *Journal of Quantitative Linguistics*, 17(2):94–100.
- Walter DAELEMANS and Antal VAN DEN BOSCH (2005), *Memory-based language processing*, Cambridge University Press, ISBN 0-521-80890-1.
- Walter DAELEMANS, Jakub ZAVREL, Ko VAN DER SLOOT, and Antal VAN DEN BOSCH (1998), TiMBL: Tilburg memory-based learner, Technical report, Universiteit van Tilburg, <https://research.tilburguniversity.edu/en/publications/timbl-tilburg-memory-based-learner-version-10-reference-guide>.
- Wolfgang U. DRESSLER (2011), The rise of complexity in inflectional morphology, *Poznań Studies in Contemporary Linguistics*, 47(2):159.

- Matthew S. DRYER and Martin HASPELMATH (2013), *The world atlas of language structures online*, Max Planck Digital Library, <https://wals.info/>.
- David EDDINGTON (2000), Analogy and the dual-route model of morphology, *Lingua*, 110(4):281–298.
- Katharina EHRET (2021), An information-theoretic view on language complexity and register variation: Compressing naturalistic corpus data, *Corpus Linguistics and Linguistic Theory*, 17(2):383–410.
- Micha ELSNER, Andrea D. SIMS, Alexander ERDMANN, Antonio HERNANDEZ, Evan JAFFE, Lifeng JIN, Martha Booker JOHNSON, Shuan KARIM, David L. KING, Luana Lamberti NUNES, *et al.* (2019), Modeling morphological learning, typology, and change: what can the neural sequence-to-sequence framework contribute?, *Journal of Language Modelling*, 7(1):53–98.
- Micha ELSNER *et al.* (2022), OSU at SigMorphon 2022: analogical inflection with rule features, in *Proceedings of the 19th SIGMORPHON workshop on computational research in phonetics, phonology, and morphology*, pp. 220–225.
- Stefano FEDERICI and Vito PIRRELLI (1997), Analogy, computation, and linguistic theory, in *New methods in language processing*, pp. 16–34, UCL Press London.
- Stefano FEDERICI, Vito PIRRELLI, and François YVON (1995a), Advances in analogy-based learning: false friends and exceptional items in pronunciation by paradigm-driven analogy, in *Proceedings of international joint conference on artificial intelligence (IJCAI'95) workshop on new approaches to learning for natural language processing, Montreal, Canada*, pp. 158–163.
- Stefano FEDERICI, Vito PIRRELLI, and François YVON (1995b), A dynamic approach to paradigm-driven analogy, in Stefan WERMTER, Ellen RILOFF, and Gabriele SCHELER, editors, *IJCAI 1995: connectionist, statistical and symbolic approaches to learning for natural language processing*, volume 1040 of *Lecture Notes in Computer Science*, pp. 385–398, Springer.
- Timothy FEIST and Enrique L. PALANCAR (2015), Oto-Manguean inflectional class database, *University of Surrey*.
- Raphael FINKEL and Gregory STUMP (2007), Principal parts and morphological typology, *Morphology*, 17:39–75.
- Joseph H. GREENBERG (1960), A quantitative approach to the morphological typology of language, *International Journal of American Linguistics*, 26(3):178–194.
- Ximena GUTIERREZ-VASQUES and Victor MIJANGOS (2018), Comparing morphological complexity of Spanish, Otomi and Nahuatl, <https://arxiv.org/abs/1808.04314>, unpublished manuscript.
- Ximena GUTIERREZ-VASQUES and Victor MIJANGOS (2019), Productivity and predictability for measuring morphological complexity, *Entropy. An International and Interdisciplinary Journal of Entropy and Information Studies*, 22(1):48.

Matías GUZMÁN NARANJO (2019a), *Analogical classification in formal grammar*, Empirically Oriented Theoretical Morphology and Syntax, Language Science Press, doi:10.5281/zenodo.3191825.

Matías GUZMÁN NARANJO (2019b), Analogy-based morphology: the Kasem number system, in Stefan MÜLLER and Petya OSENOVA, editors, *Proceedings of the 26th international conference on Head-Driven Phrase Structure Grammar*, University of Bucharest, pp. 26–41, CSLI Publications.

Matías GUZMÁN NARANJO (2020), Analogy, complexity and predictability in the Russian nominal inflection system, *Morphology*, 30:219–262.

Matías GUZMÁN NARANJO and Olivier BONAMI (2021), Overabundance and inflectional classification: quantitative evidence from Czech, *Glossa: a Journal of General Linguistics*, 6(1).

Martin HASPELMATH (2011), The indeterminacy of word segmentation and the nature of morphology and syntax, *Folia Linguistica*, 45(1):31–80.

Iván IGARTUA and Ekaitz SANTAZILIA (2018), How animacy and natural gender constrain morphological complexity: evidence from diachrony, *Open Linguistics*, 4(1):438–452.

Patrick JUOLA (1998), Measuring linguistic complexity: the morphological tier, *Journal of Quantitative Linguistics*, 5(3):206–213.

Patrick JUOLA (2008), Assessing linguistic complexity, in Matti MIESTAMO, Kaius SINNEMÄKI, and Fred KARLSSON, editors, *Language complexity: typology, contact, change*, pp. 89–108, Benjamins, Amsterdam.

Kimmo KETTUNEN (2014), Can type-token ratio be used to show morphological complexity of languages?, *Journal of Quantitative Linguistics*, 21(3):223–245.

Christo KIROV, Ryan COTTERELL, John SYLAK-GLASSMAN, Géraldine WALTHER, Ekaterina VYLOMOVA, Patrick XIA, Manaal FARUQUI, Sebastian J. MIELKE, Arya MCCARTHY, Sandra KÜBLER, *et al.* (2018), UniMorph 2.0: universal morphology, in *Proceedings of the eleventh international conference on language resources and evaluation (LREC-2018)*, European Language Resources Association (ELRA).

Alexander KOPLINIG, Peter MEYER, Sascha WOLFER, and Carolin MÜLLER-SPITZER (2017), The statistical trade-off between word order and word structure – large-scale evidence for the principle of least effort, *PLOS ONE*, 12(3):1–25, doi:10.1371/journal.pone.0173614.

Yves LEPAGE (1998), Solving analogies on words: an algorithm, in *COLING 1998 volume 1: the 17th international conference on computational linguistics*, pp. 728–735.

Yves LEPAGE (2004), Analogy and formal languages, *Electronic Notes in Theoretical Computer Science*, 53:180–191.

Vladimir I. LEVENSHTAIN (1966), Binary codes capable of correcting deletions, insertions, and reversals, *Soviet Physics Doklady*, 10(8):707–710.

Emily LINDSAY-SMITH, Matthew BAERMAN, Sacha BENIAMINE, Helen SIMS-WILLIAMS, and Erich R. ROUND (2024), Analogy in inflection, *Annual Review of Linguistics*, 10(1):211–231, ISSN 2333-9683, 2333-9691, doi:10.1146/annurev-linguistics-030521-040935, <https://www.annualreviews.org/doi/10.1146/annurev-linguistics-030521-040935>.

Gary LUPYAN and Rick DALE (2010), Language structure is partly determined by social structure, *Plos One*, 5(1):1–10, doi:10.1371/journal.pone.0008559.

Jorma LUUTONEN (1997), *The variation of morpheme order in Mari declension: suomalais-ugrilaisen seuran toimituksia*, Suomalais-Ugrilainen Seura, Helsinki.

Robert MALOUF (2017), Abstractive morphological learning with a recurrent neural network, *Morphology*, 27(4):431–458.

Stela MANOVA, Harald HAMMARSTRÖM, Itamar KASTNER, and Yining NIE (2020), What is in a morpheme? theoretical, experimental and computational approaches to the relation of meaning and form in morphology, *Word Structure*, 13(1):1–21.

Claudia MARZI (2020), Modeling word learning and processing with recurrent neural networks, *Information – an International Interdisciplinary Journal*, 11(6):320–334.

Claudia MARZI, Marcello FERRO, and Vito PIRRELLI (2019), A processing-oriented investigation of inflectional complexity, *Frontiers in Communication*, 4(48):1–23.

Clive A. MATTHEWS (2005), French gender attribution on the basis of similarity: a comparison between AM and connectionist models, *Journal of Quantitative Linguistics*, 12:262–296.

Clive A. MATTHEWS (2010), On the nature of phonological cues in the acquisition of French gender categories: evidence from instance-based learning models, *Lingua*, 120(4):879–900.

Clive A. MATTHEWS (2013), On the analogical modelling of the English past-tense: a critical assessment, *Lingua*, 133:360–373, ISSN 0024-3841, doi:10.1016/j.lingua.2013.04.002.

Peter Hugoe MATTHEWS (1972), *Inflectional morphology: a theoretical study based on aspects of Latin verb conjugation*, CUP Archive.

Matti MIESTAMO *et al.* (2008), Grammatical complexity in a cross-linguistic perspective, in Matti MIESTAMO, Kaius SINNEMÄKI, and Fred KARLSSON, editors, *Language complexity: typology, contact, change*, pp. 23–41, Benjamins, Amsterdam.

- Fermin MOSCOSO DEL PRADO (2011), The mirage of morphological complexity, in *Proceedings of the annual meeting of theGG*, 33.
- Yoon Mi OH and François PELLEGRINO (2022), Towards robust complexity indices in linguistic typology: a corpus-based assessment, *Studies in Language*, pp. 1–41.
- Enrique L. PALANCAR (2021), Paradigmatic structure in the tonal inflection of Amuzgo, *Morphology*, 31(1):45–82.
- Jeff PARKER and Andrea SIMS (2020), Irregularity, paradigmatic layers, and the complexity of inflection class systems: a study of Russian nouns, in Peter ARKADIEV and Francesco GARDANI, editors, *The complexities of morphology*, Oxford University Press, Oxford.
- Matteo PELLEGRINI and Marco PASSAROTTI (2018), Latin-flexi: an inflected lexicon of Latin verbs, in Elena CABRIO, Alessandro MAZZEI, and Fabio TAMBURINI, editors, *Proceedings of the fifth Italian conference on computational linguistics (CLiC-it 2018)*, volume 2253, pp. 324–329, Accademia University Press.
- Neil RATHI, Michael HAHN, and Richard FUTRELL (2022), Explaining patterns of fusion in morphological paradigms using the memory–surprisal tradeoff, in *Proceedings of the annual meeting of the Cognitive Science Society*.
- Benoît SAGOT and Géraldine WALTHER (2011), Non-canonical inflection: data, formalisation and complexity measures, *SFCM*, 100:23–45.
- Andrea D. SIMS and Jeff PARKER (2016), How inflection class systems work: on the informativity of implicative structure, *Word Structure*, 9(2):215–239.
- Kaius SINNEMÄKI and Francesca DI GARBO (2018), Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: a typological study of verbal and nominal complexity, *Frontiers in Psychology*, 9, ISSN 1664-1078, doi:10.3389/fpsyg.2018.01141.
- Royal SKOUSEN (1989), *Analogical modeling of language*, Kluwer Academic Publishers.
- Royal SKOUSEN (1992), *Analogy and structure*, Springer.
- Royal SKOUSEN, Deryle LONSDALE, and Dilworth B. PARKINSON (2002), *Analogical modeling: an exemplar-based approach to language*, number 10 in Cognitive Processing, John Benjamins.
- Peter SMIT, Sami VIRPIOJA, Stig-Arne GRÖNROOS, and Mikko KURIMO (2014), Morfessor 2.0: toolkit for statistical morphological segmentation, in *The 14th conference of the European chapter of the Association for Computational Linguistics (EACL)*.
- Andrew SPENCER (2012), Identifying stems, *Word Structure*, 5(1):88–108.

Nicolas STROPPIA and François YVON (2005), An analogical learner for morphological analysis, in *Proceedings of the ninth conference on computational natural language learning (CoNLL-2005)*, pp. 120–127.

Gregory T. STUMP and Rafael FINKEL (2013), *Morphological typology: from word to paradigm*, Cambridge Studies in Linguistics, Cambridge University Press.

Géraldine WALTHER and Benoît SAGOT (2011), Modélisation et implémentation de phénomènes flexionnels non-canoniques, *Traitement Automatique des Langues*, 52(2):91–122.

Robert W. YOUNG (2000), *The Navajo verb system: an overview*, University of New Mexico Press.

Matías Guzmán Naranjo

© 0000-0003-1136-6836

mguzmann89@gmail.com

University of Freiburg

Matías Guzmán Naranjo (2024), An analogical approach to the typology of inflectional complexity, *Journal of Language Modelling*, 12(2):415–475

doi <https://dx.doi.org/10.15398/jlm.v12i2.352>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

cc  <http://creativecommons.org/licenses/by/4.0/>

Corpus-based measures discriminate inflection and derivation cross-linguistically

Coleman Haley, Edoardo M. Ponti, and Sharon Goldwater
University of Edinburgh

ABSTRACT

In morphology, a distinction is commonly drawn between inflection and derivation. However, a precise definition of this distinction which reflects the way it manifests across languages remains elusive within linguistic theory, typically being based on subjective tests. In this study, we present 4 quantitative measures which use the statistics of a raw text corpus in a language to estimate to what extent a given morphological construction changes the form and distribution of lexemes. In particular, we measure both the average and the variance of this change across lexemes. Crucially, distributional information captures syntactic and semantic properties and can be operationalised by word embeddings.

Keywords:
inflection,
derivation,
morphology,
distributional
semantics,
typology

Based on a sample of 26 languages, we find that we can reconstruct $89 \pm 1\%$ of the classification of constructions into inflection and derivation in UniMorph using our 4 measures, providing large-scale cross-linguistic evidence that the concepts of inflection and derivation are associated with measurable signatures in terms of form and distribution that behave consistently across a variety of languages.

We also use our measures to identify in a quantitative way whether categories of inflection which have been considered non-canonical in the linguistic literature, such as inherent inflection or transpositions, appear so in terms of properties of their form and distribution. We find that while combining multiple measures reduces

the amount of overlap between inflectional and derivational constructions, there are still many constructions near the model's decision boundary between the two categories. This indicates a gradient, rather than categorical, distinction.

1

INTRODUCTION

In the field of morphology, a distinction is commonly drawn between inflection and derivation. This distinction is intended to capture the notion that sometimes morphological processes form a “new” word (derivation), whereas other morphological processes merely create a “form” thereof (inflection) (Booij 2007). While the theoretical underpinnings and nature of this distinction are a subject of significant and ongoing debate, it is nevertheless employed throughout theoretical linguistics (Perlmutter 1988; Anderson 1982), computational and corpus linguistics (Hacken 1994; McCarthy *et al.* 2020; Wiemerslage *et al.* 2021), and even psycholinguistics (Laudanna *et al.* 1992; MacKay 1978; Cutler 1981).

To a large degree, dictionaries and grammars roughly agree on which morphological relationships are inflectional and which are derivational within a given language. There is even a degree of cross-linguistic consistency in the constructions which are typically/traditionally considered inflections – e.g. tense marking on verbs is considered to be inflectional across a wide range of languages (Haspelmath 2024; Bybee 1985, pp. 21–22). This cross-linguistic consistency is highlighted by the development of resources such as UniMorph (Batsuren *et al.* 2022), a multilingual resource which annotates inflectional constructions across over a hundred languages using a unified feature scheme and, more recently, also includes derivational constructions from 30 languages. UniMorph data is extracted from the Wiktionary open online dictionary,¹ which organises constructions into inflections and derivations based on typical descriptive grammars for a given language, rather than any particular linguistic theory. The inflection–derivation distinction in UniMorph is therefore

¹<https://www.wiktionary.org/>

determined by what Haspelmath terms *traditional comparative concepts* (Haspelmath 2024), which are informed by the traditional structure of Western dictionaries and grammar books. The success of this initiative indicates a high degree of cross-linguistic overlap in what morphosyntactic features are considered inflectional.

Despite this relative consistency at the level of annotation, there is considerable disagreement among linguists about the fundamental properties that might underlie or explain these traditional categorisations – such as the degree of syntactic or semantic change, or the creation of new words. As an example, Plank (1994) covers no fewer than 28 tests for inflectional and derivational status. Upon applying them to just six English morphological constructions, Plank (1994) finds considerable contradictions between the results based on different criteria. Such difficulties in producing a cross-linguistically consistent definition have led many researchers to conclude that the inflection–derivation distinction is gradient rather than categorical (Bybee 1985; Spencer 2013; Copot *et al.* 2022; Dressler 1989; Štekauer 2015; Corbett 2010; Bauer 2004) or to take the even stronger position that the distinction carries no theoretical weight at all (Haspelmath 2024).

One major issue in evaluating these theoretical claims is the lack of large-scale, cross-linguistic evidence based on quantitative measures (rather than subjective tests). Work in theoretical linguistics has established that the intuitions underlying subjective tests can be problematic in certain cases (Haspelmath 2024; Plank 1994). Even so, it is possible that measures based on these subjective tests could indeed be used to classify the vast majority of morphological relationships across languages in a way that is consistent with traditional distinctions. If so, a large-scale empirical study could also provide evidence regarding the gradient versus categorical nature of the inflection–derivation distinction.

Several previous studies have shared our goal of operationalising linguistic intuitions about the inflection–derivation distinction and applying them on a large scale, but these studies have been limited in terms of both the sample size and diversity of the languages studied and the comprehensiveness and generality of the measures used. In particular, Bonami and Paperno (2018) and Copot *et al.* (2022) explored semantic and frequency-based measures of *variability* in French, aiming to test the claim that derivation tends to introduce

more *idiosyncratic* (variable) changes than inflection. Meanwhile, Rosa and Žabokrtský (2019) looked at the *magnitude* of orthographic and semantic change between morphologically related forms in Czech, following the claim that derivation tends to introduce *larger* changes than inflection. All of these studies found differences *on average* between (traditionally defined) inflectional and derivational constructions but also considerable overlap. That is, results so far are consistent with the view that although quantitative measures do align to some extent with the two traditional categories, the distinction between inflection and derivation is at best gradient. Moreover, these studies provide little evidence that quantitative measures would be sufficient to determine the inflectional versus derivational status of a new construction with any accuracy. However, it is possible that the picture could change when a wider variety of languages is included, especially if we also consider a larger number of measures at once.

In this paper, we take inspiration from both linguistic theory and the studies above to develop a set of four quantitative measures of morphological constructions, which capture *both* the magnitude and the variability of the changes introduced by each construction. Crucially, our measures can be computed directly from a linguistic corpus, allowing us to consistently operationalise them across many languages and morphological constructions. That is, given a particular morphological construction (such as “the nominative plural in German”) and examples of word pairs that illustrate that construction (e.g. “*Frau, Frauen*”, “*Kind, Kinder*”), we compute four corpus-based measures – two based on orthographic form and two based on distributional characteristics – which quantify the idea that derivations produce *larger* and *more variable* changes to words compared to inflections (Spencer 2013; Plank 1994).

We then ask whether, for a given construction, knowing just these measures is sufficient to predict its inflectional versus derivational status in UniMorph. In other words, to what extent can purely quantitative information about wordforms and corpus distribution recapitulate the linguistic intuitions, subjective tests, and comparative concepts encapsulated in the UniMorph annotations? If, across a variety of languages, belonging to different grammatical traditions, language families, and morphological typologies, the UniMorph annotations can be predicted with high accuracy based on our four measures, this would

provide evidence that traditional concepts of inflection and derivation *do* closely correspond to intuitions about the different *types* of changes inflection and derivation induce.

To explore this question, we train two different types of machine learning models (a logistic regression classifier and a multilayer perceptron). For each construction in our training set, the models are trained to predict whether the construction is inflectional or derivational, given just four input features: our measures of the magnitude and variability of the changes in wordform and distributional representations. Since we are interested in the cross-linguistic consistency of these predictors, the models are not given access to the input language or any of its typological features. In experiments on 26 languages (including five from non-Indo-European families) and 2,772 constructions, we find that both models are able to predict with high accuracy whether a held-out construction is listed as inflection or derivation in UniMorph (83% and 89%, respectively, for the two models, compared to a majority-class baseline of 57%). We additionally find that our distributional measures alone are more predictive than our formal ones, and our variability measures alone are more predictive than our magnitude ones; nevertheless, combining all four features yields the best results. Additionally, in Section 7, we investigate which *inflectional categories* are particularly likely or unlikely to be classified as inflection by our model, notably finding that inherent inflection is particularly likely to be classified as derivation by our model, in line with Booij's (1996) characterisation of inherent inflection as non-canonical.

Together, these results provide large-scale cross-linguistic evidence that despite the apparent difficulty in designing subjective tests to definitively identify inflectional versus derivational relations, the comparative concepts of inflection and derivation are nevertheless associated with distinct and measurable formal and distributional signatures that behave relatively consistently across a variety of languages. Further analysis of our results does not, however, support the view of these concepts as clearly discrete categories. Although combining multiple measures reduces the amount of overlap in feature space between inflectional and derivational constructions, we still find a gradient pattern, with many constructions near the model's decision boundary between the two categories.

In order to explore our question of interest, we need to operationalise some of the linguistic properties that have been argued to differentiate inflection from derivation. This section briefly reviews some of those properties and explains, at a high level, how they relate to corpus-based measures. We defer the detailed definitions of these measures to Section 3.

We take inspiration from the framing of Spencer (2013), who argues that morphological processes are characterised by changes to one or more of the four components of a wordform: 1. its *form* (the string of phonemes which make up its pronunciation), 2. its *semantics* 3. its *syntax* (e.g. part of speech and argument structure), and 4. its “*lexical index*”, a number corresponding to the abstract “word” to which the wordform belongs. Within this framework, a traditional view of the inflection–derivation distinction would be that inflections are those morphological relations between entries that differ in a number of aspects but have the *same* lexical index; whereas derivation corresponds to regular transformations that produce words with a *different* lexical index. Spencer argues instead for a taxonomy of morphological processes that focuses not just on lexical index, but on changes to any of these four components. Within this taxonomy, canonical inflections tend to produce small changes to one or a few components, whereas canonical derivations make large changes to more components. Indeed, in Spencer’s view, some cases classically considered derivational, such as transpositions, do not change the lexical index. Furthermore, words may be related by an inflectional process, yet (through semantic drift) have distinct lexical indices (e.g. *khaki*, a colour, and *khakis*, a type of pants). While this may seem counter-intuitive under traditional views of inflection and derivation, it is important to note that the concept of lexical index goes beyond the inflection-derivation distinction, but rather aims also to capture empirical effects observed within psycholinguistics, such as priming effects in lexical decision tasks. While it has been argued that these effects align with the inflection-derivation distinction (Laudanna *et al.* 1992; Kirkici and Clahsen 2013), this represents an independent basis for notions of words being the “same” or “different”.

While Spencer de-emphasises the classical distinction between inflection and derivation, we treat his taxonomy of morphological processes as a continuous extension of the inflection and derivation distinction. Doing so naturally unifies many existing diagnostics. It both captures and generalises correlations like derivations causing larger changes in the semantics or changing part of speech, and also suggests less frequently discussed correlations, such as derivational relations typically involving larger changes to the form of a word.² The notion of lexical index, while not directly observable, captures the notion of being the “same” or “different” word.

Importantly, it is (at least theoretically) possible to characterise a great deal of information about each of these aspects from text corpora alone. For languages with alphabetic writing systems, such as those we consider here, form is largely encoded in the orthography. Syntactic part of speech can be determined with high accuracy by the context in which words appear (He *et al.* 2018). Finally, the distributional semantic hypothesis (Harris 1954) holds that semantically similar words appear in similar types of contexts; this hypothesis is supported by the empirically impressive correlation of similarities in word embedding models like FastText (Bojanowski *et al.* 2017) with human semantic similarity judgements. However, these vectors also capture substantial amounts of information about a word’s syntactic category, as operationalised by its part of speech (Pimentel *et al.* 2020; Lin *et al.* 2015). Because of the distributional nature of meaning, it is in fact difficult to induce a space from pure language data where distance corresponds to *syntactic* similarity entirely independently from *semantic* similarity. While there is prior work on inducing such representational spaces (e.g. He *et al.* 2018; Ravfogel *et al.* 2020), due to our complex and highly multilingual setting, we instead choose to *collapse* the distinction of syntactic and semantic change made by Spencer, focusing on what is captured by embeddings designed primarily for capturing semantics but which also capture syntactic information. In particular, we use FastText embeddings, described in more detail in Section 3.2.

In addition to considering the size of the changes made to these aspects of words by a construction, we also consider the *variability*

²This is suggested, though not explicitly, by criteria like Plank’s (1994) “derivational morphemes resemble free morphs.”

of these changes. Words with different lexical indices are thought to have processes like semantic drift apply separately from each other (Spencer 2013; Copot *et al.* 2022; Bonami and Paperno 2018), which Copot *et al.* (2022) carefully links to variability in semantics. We also consider variability in the changes made to the form. This aspect has been under-explored in prior computational work. Following Plank’s (1994) claim that formal variability is greater for derivations than inflections, we would expect that allomorphy is greater for derivations than inflections, perhaps relating to the idiosyncrasies in the application of derivational allomorphs, as well as the semantic inconsistencies of derivation.

Another thread of research inspiring this particular factorisation comes from the field of natural language processing. There, the interplay between formal and distributional aspects within morphology has been widely investigated, both in derivational morphology (Cotterell and Schütze 2018; Deutsch *et al.* 2018; Hofmann *et al.* 2020), as well as in unsupervised morphological segmentation, which typically covers both inflection and derivation (Schone and Jurafsky 2000; Soricut and Och 2015; Narasimhan *et al.* 2015; Bergmanis and Goldwater 2017).

Because debates about inflectional and derivational status typically focus on *constructions* such as “the nominal plural in German” or “the addition of the *-ion* nominalisation morpheme to verbs in English,” this is the level at which we perform our analysis. Examples of constructions from our dataset are shown in Table 1. We define a construction here as a unique combination of a morpheme (given in a canonical form like *-ion* for derivation or as morpho-syntactic features for inflection), initial part-of-speech, constructed part-of-speech, and language. That is, we do not group morphemes across languages, nor do we group derivations with identical canonical forms which apply to or produce different parts of speech. This decision is motivated by examples like agentive *-er* vs. comparative *-er* in English, which differ only in the parts of speech which they apply to and produce. While there is some asymmetry in the way this grouping is handled between inflection and derivation, we do not believe this substantially affects our results. For further discussion, see Section 8.1.

Choosing to analyse constructions, rather than individual pairs of words, also has the advantage that any unusual behaviour of

Table 1: Sample of an inflectional construction (upper table, German nominative plural) and derivational construction (lower table, English verbal nominalisation with *-ion*) in our data

Base	Constructed	Morph.	Start POS	End POS	Language
Frau	Frauen	NOM;PL	N	N	DEU
Auge	Augen	NOM;PL	N	N	DEU
Lehrerin	Lehrerinnen	NOM;PL	N	N	DEU
Kind	Kinder	NOM;PL	N	N	DEU
...

Base	Constructed	Morph.	Start POS	End POS	Language
protrude	protrusion	-ion	V	N	ENG
defenestrate	defenestration	-ion	V	N	ENG
redecorate	re-decoration	-ion	V	N	ENG
elide	elision	-ion	V	N	ENG
...

individual pairs will tend to get smoothed out as we are looking at a large number of pairs for each construction (see Section 4 for details). While individual word pairs within a construction may have quite variable distributional properties, the *general tendencies* of that construction may paint a picture that is more clearly in line with notions of inflection and derivation.

Given that we are working at the level of constructions, the four quantities we wish to measure for each construction are:

- M_{Form} and V_{Form} : the average magnitude of the change in form induced by a construction, and the variability of that change.
- M_{Embed} and V_{Embed} : the average magnitude of the change in semantic/syntactic embedding space induced by a construction, and the variability of that change.

The following section describes how these measures are computed for each construction.

3

METHOD

In this section, we define M_{Form} , V_{Form} , M_{Embed} , and V_{Embed} for constructions with N pairs of words (b_i, c_i) , where b_i is the base word, and c_i the constructed word which results from applying the morphological construction.

3.1

Orthography-based measures

In this study, we use orthography as a proxy for phonological form, as discussed in Section 2. For each construction, we measure the *magnitude* of the change in form M_{Form} using the Levenshtein edit distance (Levenshtein 1966): we simply compute the average distance between each pair of words in the construction (assuming all edits count equally). For a construction with N word pairs (b_i, c_i) , this metric is given as follows:

$$(1) \quad M_{\text{Form}} = \frac{1}{N} \sum_{i=1}^N \text{EDITDISTANCE}(b_i, c_i).$$

To measure the *variability* of the change in form V_{Form} (a measure of the construction’s degree of allomorphy), we start by constructing an *edit template* for each word pair, which describes the changes made to the base in a way that abstracts away from specific string positions. For example, the pair (*tanzen*, *getanzt*) yields the edit template *ge_XXt*, meaning “start by writing *ge*, copy from the base form, delete the last two characters, and append *t*.” Similarly, the edit template for the pair (*Sohn*, *Söhne*) produces the edit template *_Xö_e*. This example highlights two important design decisions for these edit templates. First, we abstract out any variation in length of the spans which are shared with the input. This is based on the assumption that these reflect variation in the base form itself rather than morphological allomorphy. In our dataset, which does not contain any languages with templatic morphology, this assumption works well; however, future studies wishing to extend to such languages should revisit this assumption. Secondly, because we operate over orthographic form rather than the true form phonetics/featural information, edits which are considered

“the same” in linguistic theory may sometimes be considered different and vice-versa. Here, a linguist might describe this plural allomorph as adding +FRONT to the vowel’s features, which would cover the templates _Xö_e, _Xä_e, and _Xü_e. However, addressing this issue is outside the scope of this study.

Having so defined a description of the change in form with a sensible equality metric (i.e., not reliant on the length of the base), it remains to measure how much this change *varies* within a given construction. We take the edit template for each word-pair in a construction and compute its edit distance with each of the other edit templates in the construction, reporting the frequency-weighted pairwise edit distance as our measure of variability. That is, if an edit template T_i appears at a rate F_{T_i} , and there are M edit templates for a construction, this metric is computed as

$$(2) \quad V_{\text{Form}} = \sum_{i=1}^M \sum_{j=1}^M F_{T_i} \cdot F_{T_j} \cdot \text{EDITDISTANCE}(T_i, T_j).$$

For example, suppose we have a morpheme with two edit templates: _as, used 80% of the time, and _os, used 20% of the time. Then this measure would be $0.8 \cdot 0.2 \cdot \text{EDITDISTANCE}(\text{_as}, \text{_os}) + 0.2 \cdot 0.8 \cdot \text{EDITDISTANCE}(\text{_os}, \text{_as}) = 0.32$. This measure goes beyond simply counting allomorphic variants by weighting them both in terms of how different they are from each other, and by how widely they are applied in the lexicon.

Distributional-embedding-based measures

3.2

To approximate the semantic and syntactic properties of the words in our study, we use type-based (non-contextual) distributional word embeddings. Specifically, we use the FastText vectors for each language released by Bojanowski *et al.* (2017);³ these were trained on Common Crawl⁴ and Wikipedia data, which was automatically tagged by language to train language-specific embedding models (Grave *et al.*

³<https://fasttext.cc/docs/en/crawl-vectors.html>

⁴<https://commoncrawl.org/>

2018). These FastText vectors are known to correlate well with human semantic similarity scores (Vulić *et al.* 2020; Bojanowski *et al.* 2017), and are more commonly used as models of semantics than syntax.⁵ However, there is evidence from the literature in unsupervised part-of-speech tagging (He *et al.* 2018; Lin *et al.* 2015) and probing (Pimentel *et al.* 2020; Babazhanova *et al.* 2021) that they also encode syntactic information.⁶

One complicating aspect of our use of FastText vectors is that they include distributional information not only at the word, but the sub-word level. The nature of this information is itself purely distributional, relating not to the characters within those subwords, but rather the context in which the subwords appear. Nevertheless, it means that the distance between words in this distributional embedding space can be influenced by how similar they are in terms of form, when they share subwords. The primary goal of our study is identifying whether there are signals present in a raw text corpus which can reliably distinguish between inflection and derivation. As such, while the inclusion of FastText embeddings is *motivated* by their ability to represent semantic and syntactic similarity, that they include some formal information is not an issue to this primary question. It does somewhat complicate the question of assigning relative importance to formal vs distributional features, an issue we return to in Section 8.1.

⁵Recent studies have shown that embeddings from newer large language models such as mBERT (Devlin *et al.* 2019) and XLM-R (Conneau *et al.* 2020) correlate even better than FastText embeddings with human judgements of semantic similarity (Bommasani *et al.* 2020; Vulić *et al.* 2020). However, these context-dependent token-level embeddings would require further processing to produce the type-level similarities needed for our study, and we know of no strategy to do so that is validated to work with the type of resources available for our data. For example, the methods explored by Bommasani *et al.* (2020) and Vulić *et al.* (2020) are either shown to work well only for monolingual context models (which are not available for all of our languages), or only for English and multilingual models.

⁶Indeed, our own supplementary results suggests that these vectors encode substantial syntactic information, and that the addition of gold-standard syntactic category information provides little benefit over our proposed model. For further information, please see Section 2 of the supplementary material at <https://osf.io/uztgy/>.

In principle, this issue of interpretability could be avoided by using alternative embeddings that do not include sub-word distributional information, such as Word2Vec (Mikolov *et al.* 2013) or GloVe (Pennington *et al.* 2014). However, FastText has several benefits over these alternatives that we feel outweigh this issue. First, FastText models produce more accurate semantic representations of rare words (Bojanowski *et al.* 2017), which is important since many morphological variants are rare. In addition, publicly available pre-trained FastText embeddings are available for a much wider range of languages than Word2Vec or GloVe embeddings. Using these pre-trained embeddings makes our study easier to replicate and less computationally intensive, since pre-trained Word2Vec and GloVe vectors are not available for all the languages we include. It also makes our work easier to extend to other languages when relevant morphological resources become available.

Even though FastText is capable of producing vectors for words not seen at training time, we find that including these words biases low-frequency constructions to have artificially large average distances in semantic space, so we exclude all word pairs where the constructed form does not explicitly appear in the vocabulary of the FastText model. This serves as an implicit cut-off for very low-frequency forms, without requiring explicit frequency information for all of our languages.

Given the FastText embeddings, we measure changes in syntax/semantics for a construction as distances in the embedding space between the word pairs in that construction. Specifically, for each (base form, constructed form) pair (b_i, c_i) , we find the Euclidean distance between their embeddings $(E(b_i), E(c_i))$ and we compute M_{Embed} as the average Euclidean distance across all N pairs in the construction:

$$(3) \quad M_{\text{Embed}} = \frac{1}{N} \sum_{i=1}^N \|E(c_i) - E(b_i)\|.$$

While cosine distance is more frequently used than Euclidean distance for semantic similarity, this is typically because the vector norm is perceived as less relevant for semantic similarity, in part because it encodes some frequency information, at least for Word2Vec (Schakel and Wilson 2015). However, frequency information may be useful in

our case, since (as noted by Copot *et al.* 2022) the frequency of a word is correlated with the frequency of other morphological variants of that word, and more so when these variants have similar semantics. Perhaps as a result, we find this metric works as well or better than cosine distance empirically.

To measure the variability of syntactic/semantic changes within a construction, for each word pair (b_i, c_i) in the construction, we first compute the difference vector \mathbf{d}_i between the embeddings, i.e., $\mathbf{d}_i = E(b_i) - E(c_i)$. For a construction with N pairs and K dimensional embeddings, this yields a $K \times N$ matrix of differences $\mathbf{D} = [\mathbf{d}_1 \dots \mathbf{d}_N]$. We then make the simplifying assumption that the covariance between the dimensions of \mathbf{D} is zero, which allows us to estimate the variance of \mathbf{D} (and thereby V_{Embed}) as the sum of the variances of the individual dimensions k :

$$(4) \quad V_{\text{Embed}} = \sum_{k=1}^K \text{Var}(\mathbf{D}_{k,*}),$$

where $\mathbf{D}_{k,*}$ is the k -th row of \mathbf{D} .

While assuming zero covariances is not necessarily realistic (we do observe covariances which are non-zero), accurately estimating the full covariance matrix and/or its determinant requires at least as many data points as the number of dimensions in the matrix (Hu *et al.* 2017). As the number of dimensions in the FastText embeddings is 300, fulfilling such a criterion would severely limit which constructions and even languages we would be able to study here. Further, as described in Sections 5 and 6, we observe a strong empirical correlation between our measure of semantic/syntactic variability and inflectional/derivational status in UniMorph, and find this feature highly useful in creating classifiers of inflection and derivation, suggesting that this simplifying assumption does not prevent the measure from capturing relevant aspects of variability in the embedding space.

4

DATA

To perform our analysis, we require a multilingual resource that labels pairs of words with the inflectional or derivational construction that relates them. While there are many resources that provide

such construction-level information for inflectional morphology (e.g. Hathout *et al.* 2014; Ljubešić *et al.* 2016; Beniamine *et al.* 2020; Oliver *et al.* 2022), most high-quality derivational morphology resources (e.g. Kyjánek *et al.* 2020) only indicate which pairs of words are related, but not what construction relates them. An exception is the recently released UniMorph 4.0 resource, which we use in our study because it includes annotation of inflectional constructions for 182 languages as well as annotation of derivational constructions for 30 of those languages.

The data and annotations in UniMorph 4.0 are semi-automatically extracted from Wiktionary,⁷ a collection of online community-built dictionaries available for multiple languages. Inflectional and derivational information are extracted as follows:

- To identify and label inflectional constructions covering most cases, tables with the HTML class property `inflection-table` are extracted; some additional manual parsing is used to extract relations which are not tabular in some languages (e.g. English noun plurals). These tables are categorised based on their structure, and one table from each category is hand-annotated with the UniMorph feature set for inflectional features. Inflectionally related pairs, and the construction to which they belong, are then obtained from the base word associated with the entry, the particular contents of a cell, and the inflectional feature set with which that cell was annotated (McCarthy *et al.* 2020).
- To identify and label derivational constructions, the set of candidate derivations to consider for each base form A is found by looking at the *Derived terms* section of A's Wiktionary entry. The page for each derived term typically contains an etymology of the form A + -B, where -B is a derivational morpheme. In such cases, this information is added to UniMorph, together with the parts of speech of the base form and the derived term (Batsuren *et al.* 2022, 2021).

Due to the semi-automatic annotation in UniMorph 4.0, and the community-led construction of the source data in Wiktionary, there could be some errors or even systematic issues with the data. In par-

⁷<https://en.wiktionary.org/>

ticular, low-frequency forms in the inflectional data are better represented than low-frequency forms in the derivational data, because inflectional forms are constructed using paradigm tables which include all inflections of a given wordform, whereas derivational forms are added on an individual basis. However, since we necessarily exclude low-frequency forms due to the nature of our measures, this concern is somewhat mitigated. We also check for possible frequency confounds in Section 5.1.⁸

Another potential systematic issue is that the annotation may fail to collapse derivational allomorphs into a single construction. We comment further on this possible issue in Section 8.1, while noting here that our priority is to include as many languages and constructions as possible so that our sample will represent a wider range of linguistic typologies – UniMorph 4.0 contains languages with a range of morphological typologies, uncommon inflectional features, and different ratios of inflections and derivations; as well as variation in other typological variables such as syllable structure, phoneme inventory, and syntactic variables, which could affect our measures of formal or distributional change.

4.1

Data selection and summary

Of the 30 languages for which UniMorph 4.0 provides both inflectional and derivational constructions, some are not suitable for our current purposes. We exclude Galician because at time of writing its

⁸We note that data sparsity is a problem for derivational resources in general, not just UniMorph 4.0. For example, in Batsuren *et al.*'s (2021) evaluation of MorphyNet, the resource on which the derivational data in UniMorph 4.0 builds, the authors find the resource tends to have low recall and high precision when evaluated against derivational networks like Démonette (Hathout and Namer 2016), despite having comparable numbers of morphological relations. However, manual evaluation revealed that these false positives in an overwhelming majority of cases represent real morphological relationships, indicating sparsity affects both MorphyNet/UniMorph and other derivational resources. Our own manual and against-derivational-network analysis of the extended UniMorph 4.0 data showed similar trends.

UniMorph derivation data is not publicly available; Serbo-Croatian because the UniMorph data is in Latin script while the vast majority of Serbo-Croatian text used in the construction of the FastText vectors is written in Cyrillic; and Nynorsk because FastText does not distinguish between Nynorsk and Bokmål, and Bokmål is the large majority of written Norwegian.

As mentioned in Section 3.2, we exclude all word pairs where the constructed form does not explicitly appear in the vocabulary of the FastText model, due to low-quality estimates of semantic similarity for these vectors. We also exclude constructions which have fewer than 50 forms remaining after pre-processing, to ensure robust estimates of the quantities of interest. Finally, we exclude constructions where $<1\%$ of the transformed word forms are different from the base word forms, because UniMorph data is non-contextual and we would need context to distinguish the base and transformed forms. On the other hand, we ignore the problem of across-construction syncretism (where the transformed forms are identical but express different morpho-syntactic/semantic features) in the present work.

After performing the filtering steps above, we exclude Scottish Gaelic from our analysis, due to a lack of constructions that meet the inclusion criteria. This leaves us with 2,772 constructions from 26 languages: 1,587 (57.3%) of these are considered inflectional by UniMorph, and 1,185 (42.7%) are considered derivational. Table 2 contains descriptive statistics about the representation of languages, morphological typologies, and language families within our filtered dataset. Indo-European languages and, accordingly, languages with fusional typology are heavily represented in our data; however, we also have data from five languages which are not Indo-European, representing four major language families; and six languages with an agglutinative typology. We acknowledge that many language families with distinctive morphological typologies, such as the Niger-Congo languages, the Inuit-Yupik languages, and the Semitic languages, are not represented in the present study. Nevertheless, even results on a broad range of Indo-European languages plus a few others is a substantial advance in the typological coverage of existing work in the area.

Table 2: Descriptive statistics of our filtered dataset by language

Language family	Language	Morph. typology	# inf.	# der.	Total wordpairs
Indo-European (IE)	Armenian	Agglutinative	67	7	41,053
IE: Romance	Catalan	Fusional	52	31	52,329
	French	Fusional	45	104	110,643
	Italian	Fusional	50	79	127,251
	Latin	Fusional	65	23	52,175
	Portuguese	Fusional	69	35	122,622
	Romanian	Fusional	43	28	41,442
	Spanish	Fusional	121	88	337,923
IE: Germanic	Danish	Fusional	23	12	18,343
	German	Fusional	53	68	298,068
	Dutch	Fusional	21	19	36,077
	English	Fusional	7	225	119,543
	Bokmål	Fusional	14	12	50,847
	Swedish	Fusional	40	28	76,226
IE: Slavic	Czech	Fusional	96	76	103,325
	Polish	Fusional	92	104	164,837
	Russian	Fusional	94	46	292,479
	Ukrainian	Fusional	25	13	17,680
IE: Baltic	Latvian	Fusional	66	23	64,571
IE: Celtic	Irish	Fusional	21	10	21,894
IE: Hellenic	Greek	Fusional	84	3	105,358
Uralic	Finnish	Agglutinative	116	65	328,869
	Hungarian	Agglutinative	143	65	272,760
Mongolic	Mongolian	Agglutinative	16	4	15,840
Turkic	Turkish	Agglutinative	164	9	75,873
	Kazakh	Agglutinative	0	8	643
Total			1587	1185	2,948,671

DISTRIBUTION OF THE INDIVIDUAL MEASURES

In this section, we compare the distributions of our individual measures of constructions labelled as inflections to those of constructions labelled as derivations in UniMorph.

The distributions of the four measures for inflectional and derivational constructions in our data are shown in Figure 1. For all measures considered, thanks to the large amount of data in the study there is a significant difference between the mean values for inflectional and derivational constructions ($p < 0.001$ under the Mann-Whitney U test). However, we are more concerned with the direction and magnitude of those differences, which vary across the four measures.

First, looking at the form measures, we see relatively small effects of inflection-hood and derivation-hood: Cohen’s d for M_{Form} is 0.15,

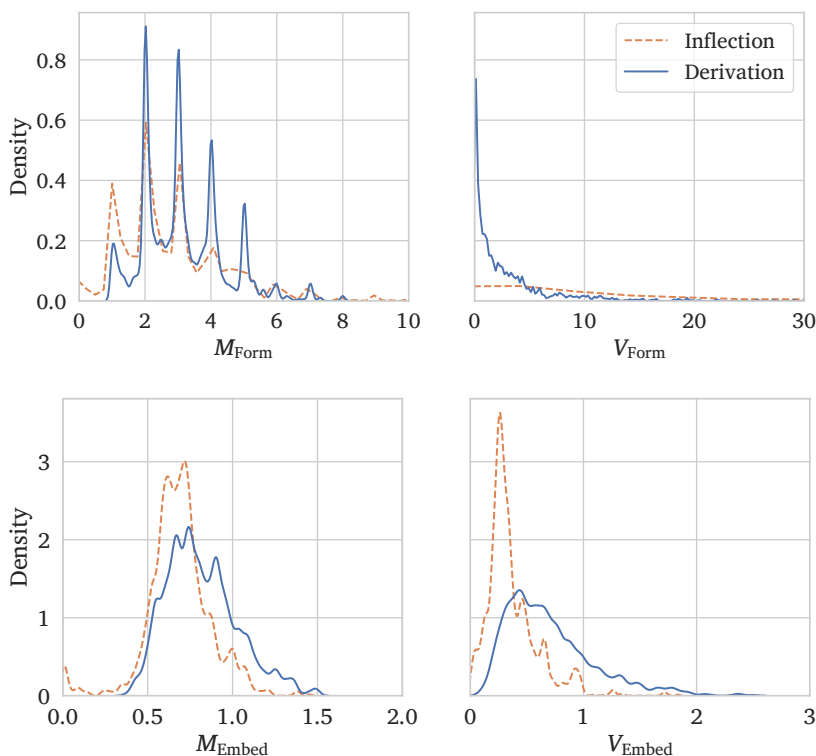


Figure 1:
The empirical distributions of our four measures (quantifying the magnitude M and variability V of changes in Form and in Embedding space) for inflections and derivations in UniMorph

while for V_{Form} it is 0.32. Despite the small difference in M_{Form} between inflection and derivation, the difference does go in the expected direction, with M_{Form} higher on average for derivation than inflection. However, on average, V_{Form} is *lower* for derivation than for inflection – the opposite of what is suggested by Plank (1994). This is discussed in Section 8.1.

In comparison to the form measures, the embedding-based semantics/syntax measures are more strongly correlated with the inflection–derivation distinction. For M_{Embed} , we observe a Cohen’s d of 0.67, indicating a moderately large effect of inflection- or derivation-hood on this measure; while for V_{Embed} we observe a Cohen’s d of 1.09, indicating a large effect. In both cases, we observe larger values on average for derivations than inflections, which indicates that relative to inflections, derivations tend to change a word’s linguistic distribution by a larger amount, and that the direction of this change is more variable. Both of these results are consistent with standard linguistic claims about inflection and derivation.

Prior work on French and Czech has suggested that any single one of these measures will show substantial overlapping regions for inflection and derivation (Bonami and Paperno 2018; Rosa and Žabokrtský 2019). Our results confirm this on a larger number of constructions and languages for all of the measures we consider.

5.1

Effects of Frequency

A potential confounder for our measures on word embeddings is frequency, since the relative frequencies of two words tend to affect their distance in distributional embedding spaces, potentially dominating or complicating meaning-related similarities (Wartena 2013). In fact, Bonami and Paperno (2018) suggested that differences in frequency may obfuscate measures of semantic distance based on current distributional embedding methods (with low-frequency constructed forms producing larger distances to a given base form than high-frequency constructed forms). If our measures are correlated with frequency, and frequency is also correlated with inflection- or derivation-hood, then any correlation we find between our measures and the inflection–derivation distinction could simply be due to this discrepancy in fre-

quency rather than to the linguistic properties of interest.⁹ Accordingly, it is desirable to quantify these relationships with frequency.

Unfortunately, for some languages considered here, word frequency information is not readily available. As a result, we restrict ourselves to the 19 languages in our data which are available through the `wordfreq` Python package. We estimate the frequency of unattested word forms as 0. We find the mean frequency of constructed inflectional word forms is less than that of derivational word forms cross-linguistically, with Cohen's $d = 0.71$, indicating a moderately large effect. However, computing Pearson's r statistic for the relationship between constructed form frequency and the four measures under consideration reveals that none of them have a significant linear association with frequency, despite the large number of word forms. While there is a sizeable relationship between some of these measures at the level of an individual distance measure (e.g. the distance between $E(\text{dog})$ and $E(\text{dogs})$), these correlations do not surface when averaged over constructions as we do in this study (e.g. the average distance between a noun and its plural form in English). As such, while our results do not contradict the concerns of Bonami and Paperno (2018), we find we are able to sidestep them in our present study by utilising a per-construction level of analysis: the effects we find here cannot be explained by frequency of constructed forms.

PREDICTING INFLECTION AND DERIVATION

6

In this section, we investigate how well the characterisation of inflection and derivation given by the UniMorph dataset can be captured by our measures. To do so, we use these measures as input features to simple classification models, which are trained to predict whether a given construction is listed as inflection or derivation in UniMorph,

⁹The reverse could also be a problem: that is, if our measures are correlated with frequency, but inflection and derivation are *not* correlated with frequency, then frequency would introduce an irrelevant confound into our measures and weaken their statistical power.

based only on those features. We created a train-validation-test split, randomly selecting 10% of the constructions to reserve for validation and 20% of the constructions for test. We used the validation set for model selection and hyper-parameter tuning, and the test set was used exclusively for evaluation of the model accuracy. We use the best model trained on this split for the analyses in Section 7 and Section 8.2. Within the current section, we evaluate our classification methods using stratified 5-fold cross-validation, to ensure the robustness of our findings to dataset splits.

To understand the scenario in which these classifiers are operating, it is helpful to consider some simple baselines. First, we note that simply predicting the majority class across languages, inflection, achieves a cross-validation accuracy of 57%, as there are simply more inflectional constructions than derivational ones in the UniMorph data. However, languages have a highly variable ratio of inflection to derivation constructions in UniMorph; classifying all the morphemes in a given *language* with the majority class for the language instead achieves an accuracy of $69 \pm 1\%$. In other words, a model could capture up to, but no more than, $\approx 70\%$ of the variation in the UniMorph data purely by capturing which language a construction is in – without achieving any ability to distinguish between inflections and derivations within a language. Note, however, that our models must predict whether a construction is inflectional or derivational without access to the language that construction comes from, so even reaching an accuracy of 70% would indicate that the input features encode cross-linguistically informative distinctions.

We tested all possible combinations of features for each of our classification models, but we focus our discussion mainly on combinations corresponding to clear hypotheses about the factors that characterise inflection- and derivation-hood. First, we consider how much any **single** feature recovers the distinction from UniMorph. Secondly, we consider several combinations of two features: (A) **just variability** ($V_{\text{Form}}, V_{\text{Embed}}$): Perhaps it is the case that only variability matters, as investigated in the embedding case by Bonami and Paperno (2018). Or perhaps (B) **just magnitude** ($M_{\text{Form}}, M_{\text{Embed}}$): only the magnitude of the changes in the components of the lexical entry matters, and variability is in practice a weak correlate or essentially redundant with magnitude. Further, it could be the case

that the two measures of either (C) **form** ($M_{\text{Form}}, V_{\text{Form}}$) or (D) **syntax/semantics** ($M_{\text{Embed}}, V_{\text{Embed}}$) alone can recover as much information as all the metrics combined. Finally, of course, there is the hypothesis (E) that **all four features** are important – each contributing some amount of unique information for recovering the distinction from Uni-Morph.

We explored these features with two types of models: a simple logistic regression classifier, which captures only linear relationships, and a multi-layer perceptron (MLP), which can capture non-linear relationships between features. The logistic regression classifier encodes the assumption that inflection and derivation can be separated by a hyperplane in feature space. If the feature values cluster, without intermediate regions, this corresponds to a categorical characterisation of the distinction. If there are instead large regions with intermediate values, this corresponds to a gradient characterisation of the distinction.¹⁰ If the non-linear model is required to recover the distinction, then discontinuous areas in the feature space may fall in a certain category, which would not neatly correspond with linguistic intuitions.

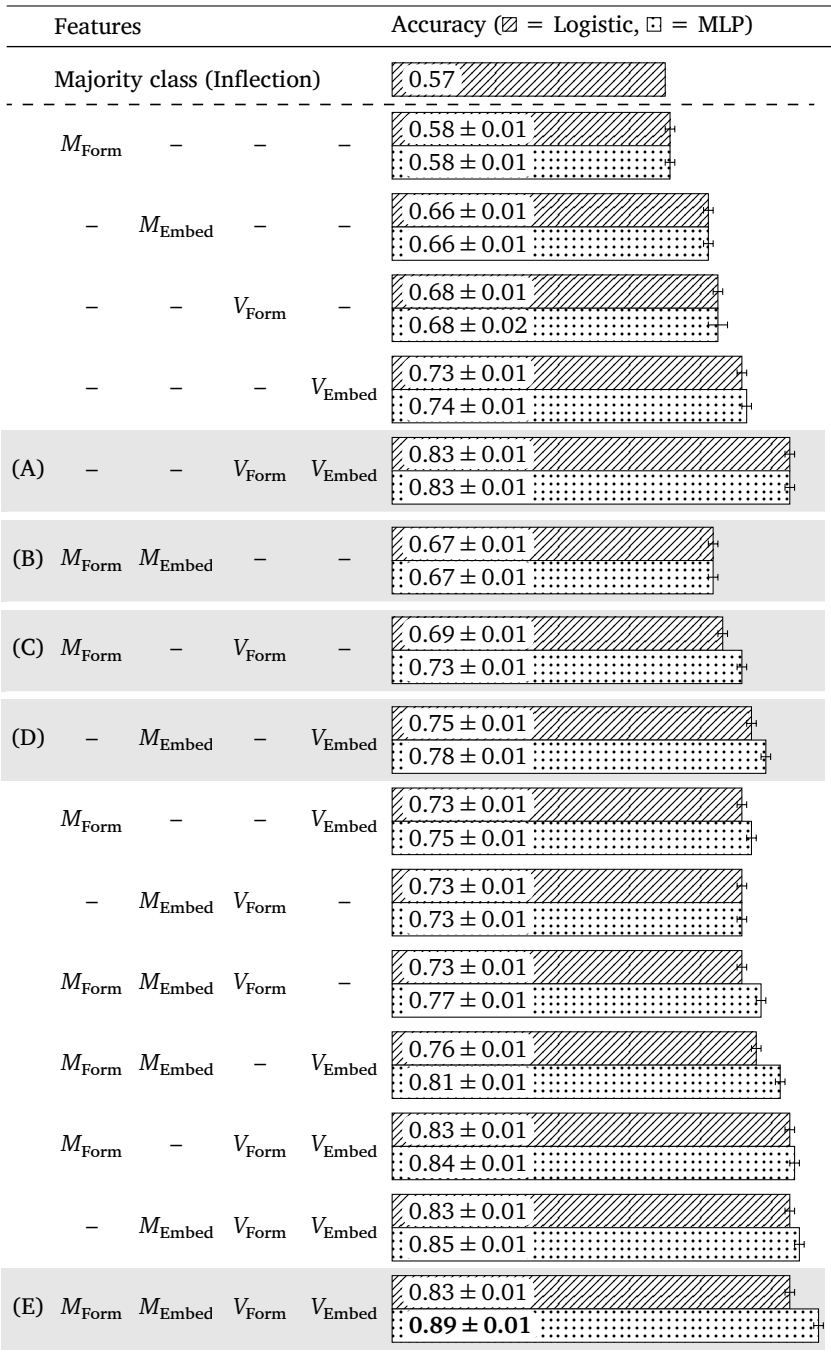
First, we consider the logistic regression classifier. As described in Section 2, the expectation from linguistic theory is that greater values of any measure should be associated with that construction being derivational. Our analysis in Section 5 largely backs up this relation (with the relationship being inverted for form variability), though it is not clear to what degree this relationship is strictly linear.

Due to our highly-restricted selection of measures, we are able to create classifiers with all possible combinations of features. As shown in Figure 2, the logistic classifier results best support the **just variability** hypothesis (A), with no notable performance gains achieved by adding other features in a linear-modelling setting.

While our best logistic classification model can capture 26 points of variation more than predicting the majority class, it may be missing non-linear interactions between independent variables, or between an individual independent variable and the dependent variable. To account for such non-linear relationships, we fit a multi-layer perceptron (MLP) with a hidden layer size of 100, using the Adam optimiser

¹⁰ This issue of whether the distinction is gradient or categorical with respect to our measures is discussed further in Section 8.4.

Figure 2:
Cross-validation
accuracy and
standard error
in reconstructing
UniMorph’s
inflection-
derivation
distinction
by various
supervised
classifiers.
Linguistically-
motivated
hypotheses
referred to in the
text are denoted
with letters



(Kingma and Ba 2015) and training for 3000 steps. The number of layers and layer size was chosen using validation set performance, while the number of steps was chosen based on loss convergence on the training set. We find similar patterns of performance for most combinations of predictors. However, we see substantial improvements in performance for combinations of features which include both magnitude and variability features; for example, $(M_{\text{Form}}, V_{\text{Form}})$ improving from $69 \pm 1\%$ to $73 \pm 1\%$. Perhaps as a result of this, we achieve a test-set accuracy of $89 \pm 1\%$, when using all four predictors – representing a 6-point improvement over the best linear model, as well as a 4-point improvement over the best combination of three measures using the MLP $(M_{\text{Embed}}, V_{\text{Embed}}, V_{\text{Form}})$. This therefore suggests that while the variability features are the most descriptive of UniMorph’s categorisation of inflection/derivation, all four features contain unique information relevant to recreating this distinction (Hypothesis E).

CLASSIFICATION OF LINGUISTIC TYPES OF INFLECTION

7

Given the controversy over what should be considered inflection and derivation, a model that largely aligns with a typical operationalisation of the distinction (UniMorph 4.0) may also be of interest in the ways in which it *differs* from that operationalisation. Accordingly, in this section, we look at the trends in how our model classifies constructions which are labelled as inflection in UniMorph. We consider several distinctions which we believe to be of linguistic interest, specifically: what kind of meaning is expressed by an inflection; whether it is *transpositional* (changes the part of speech); and whether it is *contextual* or *inherent* (as described by Booij 1996). We ask whether these distinctions affect how likely an inflectional construction is to be classified correctly under our best model (the MLP with all four measures). We focus only on inflectional constructions because UniMorph has cross-linguistically consistent featural annotations on inflections that we can use for the analysis; no such cross-linguistically consistent annotation exists for derivation.

7.1

Categories of inflectional meaning

We first consider several categories of inflectional meanings: features for mood (e.g. indicative, subjunctive); tense (present, past...); number (singular, dual, plural ...); voice (active, passive); comparison (comparative, absolute/relative superlative, equative); gender, and case. These categories of meaning are often used to structure accounts of inflection, such as UniMorph's description of its feature set (Sylak-Glassman 2016) as well as theoretical accounts like Anderson (1985) and even Haspelmath's (2024) retro-definition of inflection. It is, however, worth noting that not all sources agree on all of these categories as being inflectional. For example, Haspelmath rejects voice as inflectional, and comparison is often omitted from discussions of major cross-linguistic inflectional categories (as is the case in both Anderson 1985 and even Haspelmath 2024), and is considered *inherent inflection* (which is less canonical) by Booij (1996). One might reasonably expect constructions which are semantically marked for these controversial categories to be *more likely to be classified as derivation* by our model.

Note that linguists generally agree on which categories of meaning are semantically marked across languages (Greenberg 1966; Silverstein 1986; Croft 2002; Ackema and Neeleman 2019), and semantic markedness often corresponds to morphological marking. For example, past tense is generally considered more semantically marked than present, and in many languages the past tense requires an affix while the present tense does not. However, the UniMorph annotations include both the semantically marked and unmarked inflections (e.g. V;PAST;PL and V;PST;PL for Ukrainian verbs). Therefore, for the purposes of this analysis, we consider active voice, singular number, nominative case,¹¹ and present tense unmarked values, even when present in the featural description of a construction. For example, in Ukrainian verb annotations, V;PAST;PL would be considered marked for tense and number, while V;PST;SG would be considered unmarked for both; both verbs would be unmarked for voice and mood since

¹¹ While some languages have been argued to mark for nominative case with accusative being unmarked (König 2006), no such language is present in our study.

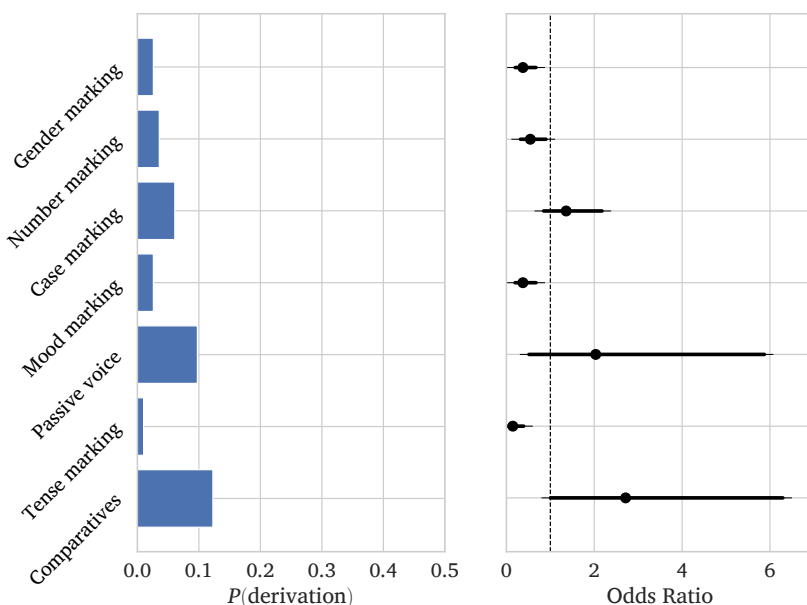


Figure 3: Probability and Odds ratio with 95% confidence intervals of being classified as derivation for various kinds of inflectional meaning. Inflections to the right of the dotted line were disproportionately likely to be classified as derivation by our model

these are not in the featural descriptions. For the category of gender, we simply consider nouns not to be marked, as their gender is typically not a morphological process but a lexical property.

Figure 3 displays the probability that a construction marking for one of these inflection types will be classified as derivation by our best-performing model. As can be seen in the figure, our model does not classify any of these major kinds of inflection as *more derivational than inflectional*; each is substantially more likely to be classified as inflection than derivation. This finding is perhaps unsurprising given our model’s cross-linguistic test set classification accuracy of 90% – it classifies 92% of inflections correctly in general. Accordingly, classifying just 15–20% of constructions belonging to a particular inflectional category as derivations has the potential to be significant.

In order to answer the question “Are constructions which mark for this inflection type significantly more likely to be classified as derivational than others?”, we compute the odds ratio. We focus on the best

performing MLP model (using all 4 features) in these results, which are presented in Figure 3 with 95% confidence intervals. Constructions with an odds ratio significantly greater than 1, while not more likely to be classified as derivation than inflection, can nevertheless be thought of as particularly *non-canonical* types of inflection under our model, while those with odds ratios significantly below 1 are *canonical* with respect to our model.

We apply the Boschloo exact test (Boschloo 1970) to the results and correct for multiple comparisons with the Bonferroni correction, which yields a significance level of $0.05/7 = 0.007$. We find the odds ratios for gender ($p = 1 \times 10^{-7}$), tense ($p = 3 \times 10^{-7}$), and mood ($p = 1 \times 10^{-7}$) significant. This identifies gender, mood, and tense as particularly canonical inflectional distinctions under our model – all of which are well in line with the claims of Haspelmath and others.

While we do not identify any inflectional meaning categories which are significantly more likely to be classified as derivations than the average inflections, the categories of passive voice ($p = 0.03$) and comparatives ($p = 0.08$) each have 95% confidence intervals which are almost exclusively larger than 1. Each of these categories has been discussed as less canonical kinds of inflection, with comparatives even occasionally being listed as derivations within UniMorph.¹² As these are the two least common categories in our sample (consisting of just 57 comparative constructions and 41 passives), it may be that these effects would be significant with a larger sample; alternatively, their relatively high likelihood of being classified as derivation could be an artefact of their rarity in our sample.

7.2 *Inherent vs. contextual inflection and transpositions*

While we do not find any categories of inflectional *meaning* as non-canonical under our model, we also consider two other major categories of inflection that have been discussed in the linguistic literature as potentially non-canonical: inherent inflection and transpositions, for which results are displayed in Figure 4.

¹²For example, they are listed as derivations in English, but as inflections in German.

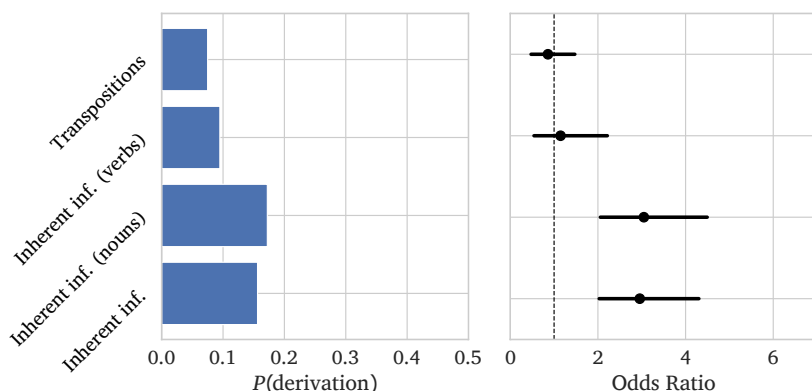
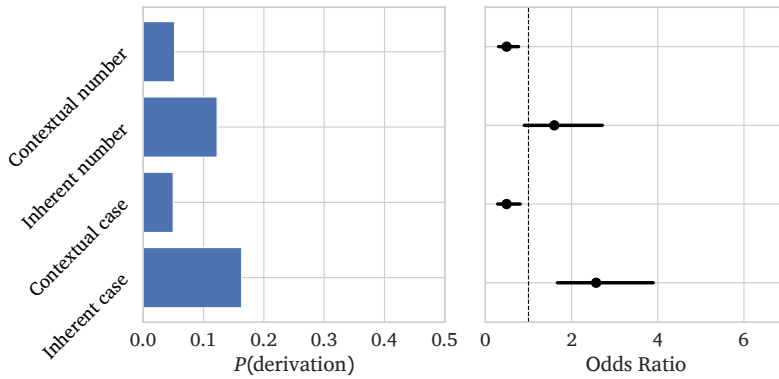


Figure 4:
Probability
and Odds ratio
with 95%
confidence
intervals of being
classified
as derivation
for inherent
inflections and
transpositions

First, we consider Booij's (1996) notion of inherent and contextual inflection. Booij describes contextual inflection as canonical: it is determined by the syntactic context in which a word appears and indicates agreement (e.g. plural marking on a verb, which is controlled by its subject). In contrast, inherent inflection is non-canonical: it contributes to the meaning of the word itself (e.g. the plural noun). To operationalise this in a simple, cross-linguistically consistent way, we associate number, gender, and case¹³ with nouns – meaning that when those features appear on other parts of speech, we consider them contextual inflections. Analogously, we associate mood, tense, and voice with verbs. We then may consider whether an inflection is *inherent* or not, where we define inherency as not marking *any* contextual features. As shown in Figure 4, we find that inherent inflectional constructions are not more likely to be classified as derivation than inflection; however, they *are* significantly more likely to be classified as derivation compared to other types of inflections, as quantified by the odds ratio ($p = 6 \times 10^{-9}$). Interestingly, though, we find this to be almost entirely due to nominal inherent inflection ($p = 2 \times 10^{-8}$), rather than verbal inherent inflection ($p = 0.7$). We see this exemplified in Figure 5, which shows that inherent case is significantly associated with being classified as derivation ($p = 1 \times 10^{-5}$), while contextual

¹³ Booij (1996) makes the distinction between structural and semantic case, with the former being contextual inflection and the latter inherent. However, due to the complexity in drawing a line between these categories, we treat all case marking on nouns as inherent.

Figure 5:
Probability and
Odds ratio with
95% confidence
intervals of being
classified
as derivation
for inherent vs.
contextual noun
inflections



case ($p = 0.003$) and contextual number ($p = 0.0008$) are significantly associated with being classified as inflection.

Finally, we consider inflectional transpositions, denoted in UniMorph as participles (deverbal adjectives), converbs (deverbal adverbs), and masdars (deverbal nouns), shown in Figure 4. Transpositions have often been argued to be non-canonical inflection or even derivation because transpositions change the part of speech (Spencer 2013; Plank 1994; Haspelmath 2024). We here find under our model that transpositions appear neither significantly more or less likely to be classified as derivations than inflections by our model – neither particularly canonical or non-canonical. This may be due to the non-contextual nature of our embedding model: many inflectional transpositions are syncretic with a non-transpositional form, and our model must assign these the same location in embedding space. Thus, our null result here should not be taken as strong evidence against considering transpositions as non-canonical.

7.3

Summary

In this section, we have investigated different kinds of inflectional constructions discussed in the linguistics literature to see whether any of these are particularly *canonical* or *non-canonical* under our model. That is, we looked at whether our model is more (or less) likely to correctly classify these constructions as inflectional, relative to the average inflectional construction.

We identify mood, tense, and gender as *canonical inflections* under our model, but we do not find any categories of inflectional meaning which are significantly *non-canonical* in our sample. We find that inherent inflections are significantly more likely to be classified as derivations, in line with Booij’s (1996) view of them as non-canonical inflection. Interestingly, we find this is driven by inherent nominal inflections rather than inherent verbal inflections. Finally, we investigate transpositions (typically thought of as non-canonical inflection), finding no evidence that they are either canonical or non-canonical under our model.

DISCUSSION

8

The role of our individual measures

8.1

As shown in Section 6, all four of our measures can be used to achieve better discrimination between traditional concepts of inflection and derivation; however, not every feature plays an equally large role. In this section, we discuss the roles played by each of our features and their connection to linguistic theory.

Among our four measures, our results point to variability of the change in distributional embedding V_{Embed} being the most relevant to traditional categorisations of inflection and derivation. This is in line with the findings of Bonami and Paperno (2018) and Copot *et al.* (2022) in French, who focus on similar measures as a proxy for semantic drift, as part of a theory where traditional concepts of inflection and derivation reflect higher or lower *paradigmatic predictability*. Indeed, it is possible that this measure could be (roughly) equivalent to Copot *et al.*’s (2022) predictability of frequency, as it is motivated from a similar theoretical basis. On the other hand, our measure is much simpler to define and compute: attempting to produce a measure of *predictability* immediately raises complex issues around on *what basis* such predictions should be made, complicating the interpretation of results.

In addition, we find a clear and complementary influence of the variability of the change in form, V_{Form} : adding this feature to our

model produces a large increase in performance, even when V_{Embed} is already included. This measure (described in Section 3.1) can be thought of as a weighted measure of allomorphy, capturing not just the number of distinct patterns, but also their similarity. Our results point to a much higher degree of formal variability/allomorphy for inflections than derivations across a wide range of languages, contrary to the predictions of Plank (1994) and Dressler (1989). Although work on French has suggested little difference in the *predictability* of form for derivational and inflectional constructions (Bonami and Strnadová 2019), we clearly find within our sample of languages evidence that the *actual degree of variation* is very different.

Superficially, this finding could appear to be caused by the fact that derivational allomorphs are sometimes not collapsed in UniMorph data (e.g. *-heit* and *-keit* being listed as different morphemes in German). However, when we looked into this issue, we found that most derivations had 0–1 such uncollapsed allomorphs. Combining two allomorphs in this way would add at most half the edit distance between the morphs to our measure. In most cases, the edit distance between these allomorphs is 1–2, adding just 0.5–1.0 to the value of V_{Form} . This is much less than the difference between the means of the two categories in this feature, suggesting that failure to collapse allomorphs is not the primary source of this finding. Returning to the example of *-heit* and *-keit* within German, we find *-heit* has V_{Form} of 1.53 and *-keit* has V_{Form} of 1.25. The two morphemes occur 27% and 73% of the time respectively. When combined, they have a V_{Form} of 2.43 – still well within the derivational range.

Similarly, one might object that not only such straightforwardly-conditioned allomorphs must be accounted for, but also more idiosyncratic variants that express the same meanings. For example, in French, such formally distinct forms as *-age*, *-ance*, and *-ure* could be argued to be allomorphs of a single action-noun forming morpheme. Copot et al. (2022) handle this by grouping morphemes with similar semantics, by computing average difference vectors in embedding space between base and constructed form for each morpheme, and agglomeratively clustering morphemes with difference vectors with cosine similarity over 0.7. We find such clustering of our data does not sufficiently align with semantic categories of morphemes across

our full range of languages to reformat our analysis around it. However, even when clustering derivations with this threshold of similarity, we still find a much lower degree of formal variability for derivations than inflections. On average across languages, 38% of derivational constructions cluster with nothing else at all, without increasing variability. The average cluster contains just 1.8 morphemes, with inflectional morphemes, which are not clustered in this way, exhibiting still 208% more allomorphs on average than derivational clusters.

Future studies should explore the relevance of the variability of form further, to see if it is robust to different languages, and focus directly on the validity of this measure. However, we note that our best performing model without this feature, the MLP with the features $(M_{\text{Form}}, M_{\text{Embed}}, V_{\text{Embed}})$ achieves a classification accuracy of $81 \pm 1\%$, which is still 23 points above predicting the majority class.

Finally, our results show smaller influence of the magnitude measures M_{Form} and M_{Embed} . This finding seems to contrast with Spencer's general claim that derivations are associated with larger changes to the properties of a lexeme, but it is not entirely contradictory. In particular, M_{Embed} still displays a fairly strong correlation with inflection and derivation on its own, and likely does not contribute as much to our models due to its substantial correlation (Pearson's r : 0.86) with the more strongly predictive V_{Embed} . In the case of M_{Form} , we find little evidence here that derivations have a tendency to produce larger changes to the form; however, this may be in part related to our need to remove constructions which are orthographically syncretic between the base form and constructed form (which are dominantly considered inflectional in our sample of languages). The length of the change in form does seem to play a small role as a part of a composite set of factors based on its use in our best-performing MLP model.

As noted in Section 3.2, our use of FastText somewhat complicates the interpretation of the role of the distributional measures, in the sense that embeddings based on sub-words may capture some formal similarity between words as well as semantic and syntactic similarity. However, we note that if the embeddings do capture formal similarity, at least some of this information must be complementary to that captured by our form-based measures, since including both

types of features yields a better classifier than either alone. We also performed some supplementary experiments with Word2Vec embeddings to check that distributional features without sub-word information are also useful.¹⁴ While overall performance of the classifier was lower (likely due to overall worse quality of the embeddings, for the reasons described in Section 3.2), we still found a non-trivial contribution from the distributional features. So, while we can say that both formal and distributional properties are associated with the inflection-derivation distinction, further work is needed to clearly distinguish semantic, syntactic, and formal properties.

8.2

Language generality

An important aspect of our model is its language-generality. A major limitation of existing computational studies of the inflection-derivation distinction (Copot *et al.* 2022; Rosa and Žabokrtský 2019; Bonami and Paperno 2018) is their focus on single European languages. In particular, Haspelmath (2024) argues that many properties of inflection and derivation are not proven to apply in a consistent way across languages (especially non-European and non-Indo-European languages). Our model achieves high accuracy across languages, while using no language-specific features. As such, it suggests that across the languages in our sample, inflection and derivation show cross-linguistically similar distributional properties.

Given the large number of European languages in our sample, this result clearly suggests that, at least in the Indo-European family, inflection and derivation are associated with distinct signatures in terms of both their distribution and their form (at least, as expressed in orthography). While evidence for such claims has been provided in specific languages by Copot *et al.* (2022), Bonami and Paperno (2018), and Rosa and Žabokrtský (2019), many large sub-families within the Indo-European language family had previously been untouched by this literature. Our study includes several Germanic languages with distinctive morphological traits, as well as Armenian, Latvian, Irish,

¹⁴For more details about these experiments, see the supplementary material at <https://osf.io/uztgy/>.

and Greek, covering many smaller European branches of the Indo-European family. We also expand the evidence for consistency in the application of the terms “inflection” and “derivation” within the Romance and Slavic language families. This broad coverage overall provides quantitative evidence for the cross-linguistically consistent application of the inflection–derivation distinction within the languages of Europe – not only in terms of the morpho-syntactic traits of these constructions, as framed by Haspelmath (2024), but also in terms of corpus-based measures which are a proxy for the linguistic intuitions and subjective tests Haspelmath argues should be abandoned.

In addition to this robust evidence that these properties can discriminate inflection and derivation within Indo-European languages, we also show evidence of a degree of applicability to a wider range of languages. On this subset of languages, our best MLP classifier averages 82% accuracy on the test set, lower than for the Indo-European languages (average 91% accuracy). While this is still well above the majority class baseline (74% accuracy on this subset), it suggests that the application of the inflection–derivation distinction to non-Indo-European languages may indeed be less consistent, as suggested by Haspelmath. Of particular note are the results for Turkish. Turkish is a highly agglutinative language with, according to traditional descriptions, an exceptionally rich inflectional system – reflected by an extremely large number of inflectional constructions and relatively small number of derivations in our dataset. Our classifier over-uses the label derivation for this language – classifying all derivations correctly, but also classifying many inflections as derivations. This suggests a mis-alignment between the orthographic and distributional tendencies observed in European languages, and the way linguists typically operationalise inflection and derivation in this language. On a theoretical level, then, our results are therefore compatible with either a view where we should think of some of these so-called inflections in Turkish as more derivational, or a view where these corpus-based measures are less accurate indicators of what “should” be considered inflection for Turkish.

Due to the relatively small number of non-Indo-European languages and constructions from these languages we are able to consider in the present work, we are unable to draw definitive general conclusions about cross-linguistic consistency in our measures with

languages outside Europe. Our results here seem to point to an intermediate view where these corpus-quantifiable correlates of inflection and derivation are *less reliable* descriptors of the way the distinction is made outside of Indo-European languages but still explain *substantial amounts* of the distinction.

8.3

The classification approach

Another key differentiating aspect of our work from previous computational studies is our focus on classification of constructions. This method allows us to quantify *how much* of the inflection–derivation distinction, as operationalised across a wide range of languages, can be explained by our simple set of corpus-based correlates. Our focus on a wide range of languages necessitates the use of a quantitative method such as classification, and contrasts with the single-language studies of Bonami and Paperno (2018) or Copot *et al.* (2022), who focus more on discussing individual constructions.

Further, our goal of looking at whether *multiple features* produces a more clear-cut and less gradient view of inflection compared to the single correlates examined by Bonami and Paperno (2018) or Copot *et al.* (2022) prevents us from simply doing a statistical test of correlation between a feature and inflection/derivation. While we avoid this by training a classification model, Rosa and Žabokrtský (2019) solve this problem by using clustering. We believe doing so conflates two questions about the measures under consideration. First is the question of how *consistent* linguists’ categorisations are in terms of the measures. Secondly, there is the question of how *natural* the traditional categories of inflection and derivation appear with respect to these measures. This first question is a lower bar than the latter: it may be possible to use these measures to determine inflectional or derivational status, regardless of whether they form natural clusters in the feature space.

Nevertheless, a finding of *consistency* without *naturalness* is still interesting, given that decisions about what to consider inflection and derivation were made without access to these measures. For example, consistency with respect to these measures could make them a successful “retro-definition” in the terms of Haspelmath (2024). The clustering approach may also fail to identify a distinction where inflection

and derivation are predominately located in only slightly overlapping regions of the feature space but do not necessarily form natural clusters.¹⁵ It is this question of consistency which we primarily consider in this paper, leading us to eschew the unsupervised clustering approach for supervised classification.

Another advantage of our focus on classification is that it naturally lends itself to testing the *generalisability* of our claims: by holding out a random subset of our constructions for testing data and computing accuracy on that set, we confirm that our results do not over-fit to the constructions in the training set.

Inflection and derivation: gradient or categorical?

8.4

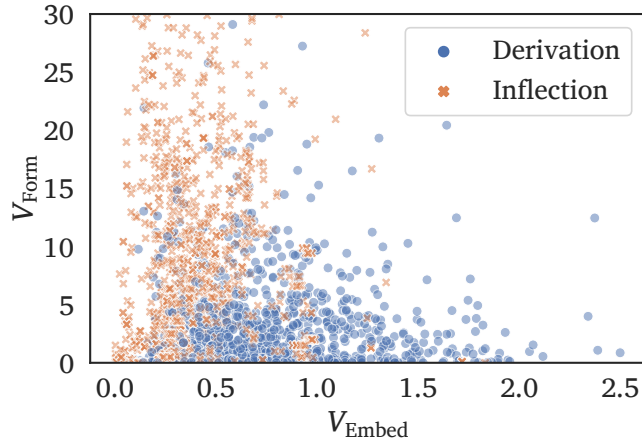
Whether the inflection–derivation distinction is principally a gradient or categorical phenomenon is a longstanding debate within linguistic theory with potentially wide-ranging implications about the nature of linguistic representations. Many theories of morphological grammatical organisation, production, and processing implicitly or explicitly employ the “split morphology hypothesis,” which holds that inflection and derivation are separated in the grammar (Perlmutter 1988; Anderson 1982). Those who propose such separate structures rely on both the distinction between inflection and derivation being discrete and the specifics of that distinction – i.e., what morphological constructions in what languages are considered either inflectional or derivational.

On the other hand, a growing body of linguistic theory rejects a hard distinction (e.g. Bybee 1985; Spencer 2013; Dressler 1989; Štekauer 2015; Corbett 2010; Bauer 2004). In its place, they often treat inflection and derivation as a gradient, perhaps emergent out of deeper phenomena. This view has been borne out in the computational work of Bonami and Paperno (2018) and Copot *et al.* (2022) who find clear continuous gradience with respect to their metrics and the categories of inflection and derivation.

While, as discussed in Section 8.3, we focus primarily on the *consistency* of traditional categories of inflection and derivation, in this

¹⁵ As described in Section 8.4 and shown in Figure 6, it is this situation in which we find ourselves.

Figure 6:
Our two most predictive
measures for inflection
and derivation. Saturation
represents overlapping
constructions. With respect
to these two variables,
the inflection–derivation
distinction appears
gradient rather than
categorical



section we briefly investigate whether, under our measures, the distinction between inflection and derivation appears more *gradient* or more *categorical*. If the former is the case, we expect a relatively even distribution of constructions in feature space, which (perhaps gradually) transition from being traditionally classified as inflection to being traditionally classified as derivation. In the categorical case, however, we expect *clusters* within feature space with relatively few constructions lying in intermediate ambiguous regions.

We focus on four measures in this study, so we are unable to directly visualise in the feature space. While we applied principal component analysis to produce a two-dimensional representation of our full feature space, the principle components did not pattern into inflectional and derivational regions. This is certainly evidence against *naturalness* of the traditional distinction with respect to our measures. However, we may also look at our two most strongly predictive measures, as shown in Figure 6. Recall that a logistic classifier using only these features was able to correctly classify $83 \pm 1\%$ of constructions. Our results with our measures are here consistent with the existing findings of a gradient, rather than categorical, distinction between inflection and derivation with respect to traditional linguistic tests/measures which operationalise them – we observe a spread of constructions in the two-dimensional feature space with a smooth transition between regions containing almost exclusively inflections and regions containing almost exclusively derivations.

*Are inflection and derivation identifiable
from the statistics of language?*

8.5

In this work, we have focused on identifying cross-linguistically applicable corpus-based measures, which have a consistent relationship with the traditional concepts of inflection and derivation. While we have primarily motivated the use of these corpus-based measures in terms of quantifying how consistently these categories are applied across languages or making concrete subjective linguistic tests, the fact that they are built purely from the statistics of natural language corpora allows us to consider another important question: is the inflection-derivation distinction something which is present in the statistics of language itself?

If the retro-definition given by Haspelmath (2024) is the right one, for instance, the answer to this question would superficially appear to be *no*. Haspelmath casts the distinction in terms of morpho-syntactic feature values, which themselves refer in many cases to the *meaning* expressed by a morphological exponent. If the specific meaning expressed by a morphological relation is necessary to distinguish which relations are inflectional in nature and which are derivational, then the typical inflection–derivation distinction requires *grounding* the meanings of sentences to solve – for example, no amount of raw text input in a language can tell you whether the relationship between two words is “agentive” or “plural.”

The answer to this question has implications within psycholinguistics as well as computational linguistics. Psycholinguistics provides some empirical evidence that inflection and derivation are processed differently (Laudanna *et al.* 1992; Kirkici and Clahsen 2013), which seems to imply learners have some implicit ability to categorise constructions into inflection and derivation. How might a learner learn what processing to apply to a given morphological construction in this case? A substantial body of literature indicates that humans can and do perform purely statistical learning within language acquisition (Swingley 2005; Saffran *et al.* 1996; Thiessen *et al.* 2013; Thompson and Newport 2007; Thiessen and Saffran 2003). Without using or even having access to the references of sentences in some cases, learners uncover important aspects of the structure of language. Our results therefore suggest the possibility that statistical learning may

play a role in learning to process canonical inflection differently from canonical derivation.

This is also relevant for the validity of several constructs within natural language processing. For example, the paradigm clustering task from SIGMORPHON 2021 (Wiemerslage *et al.* 2021), which requires identifying inflectional paradigms from raw text, can only be solved if inflections and derivations can be distinguished from the statistics of such a corpus. Otherwise, derivational relations would be outputted by even the best possible system. Similarly, the task of unsupervised lemmatisation (Kasthuri *et al.* 2017; Rosa and Zabokrtský 2019) also relies on the distinction between inflection and derivation being evident within a text corpus. Our results point to these types of construct being largely valid for Indo-European languages given the high degree of discriminability between the categories, but our slightly lower results for non-Indo-European languages suggests the need for further investigation into the validity of such constructs for typologically-distant languages to those considered here.

8.6

Future work

We believe our study presents a number of interesting avenues for expansion. One such possibility is the extension of the present work to a larger and more diverse sample of languages. In this work, we have taken advantage of the recently produced UniMorph 4.0 dataset to validate claims based on individual languages that corpus-based measures can capture traditional notions of inflection and derivation, and quantify how many intermediate constructions exist under such measures, but our results mostly bear on languages of Europe belonging to the Indo-European language family. While this still represents a substantial advancement in knowledge, and we do find some evidence that our results are applicable to non-Indo-European languages (as described in Section 8.2), the evidence presented here cannot yet fully refute Haspelmath’s (2024) claim that inflection and derivation are much less applicable to languages outside Europe. Relatively few (590) of the constructions in our data belong to non-Indo-European languages, with even fewer (201) coming from languages spoken outside Europe, and no representation of languages from outside Eurasia. As argued by Dryer (1989), typological claims must be made not

just with normalisation with respect to language families or small geographical areas, but even large geographical areas – which is not possible with available data. In order to properly understand to what degree the concepts of inflection and derivation map onto language generally, there is a critical need for the expansion of resources like UniMorph 4.0 and Universal Derivations (Kyjánek *et al.* 2020) to cover a larger and more representative set of languages. While UniMorph increasingly covers the inflectional morphology of a wide range of languages throughout the world, having added 65 languages from 9 non-European language families in the 4.0 release alone, no unified derivational resource covers a large number of non-European languages. The harmonisation and integration of resources like derivational networks such as Hebrewnette (Laks and Namer 2022) and finite-state morphological transducers which cover derivation such as Arppe *et al.* (2014–2019), Larasati *et al.* (2011), Strunk (2020), or Vilca *et al.* (2012) into multilingual resources is essential to answering truly general typological questions with these resources in the future.

Another limitation of this study that future work could address is indeed our use of the UniMorph 4.0 dataset. While UniMorph 4.0 provides the largest-scale multilingual dataset of inflection and derivation presently available, it is limited by factors related to its semi-automated construction, which may affect the way allomorphy is represented (as discussed in Section 8.1), or other as-of-yet undiscovered systematic biases.¹⁶

Additionally, we have limited ourselves to a small set of measures here. Future work could seek to improve these measures, or look at other or additional measures. Many previously suggested properties of these categories, such as affix ordering, have directly observable effects on the statistics of text. Future works could test corpus-based measures of distance from the stem or limitedness of applicability, for

¹⁶ See Malouf *et al.* (2020) for a discussion of potential pitfalls of the UniMorph dataset for typological research. UniMorph represents not exactly a consensus of highly-trained linguists, but rather largely of the amateur lexicographers that make up the Wiktionary community. Accordingly, as more large-scale multilingual datasets are available, future work should investigate the degree to which these findings are robust to the method of data collection as well as the source of the data.

example. Particularly interesting, we believe, would be the investigation of a syntactic distance and variability component, drawing on works such as He *et al.* (2018) and Ravfogel *et al.* (2020) – though there are significant challenges to operationalising these embeddings in a multilingual, low-resource domain.

There is also room for refinement of our measures and classification techniques. For example, extension to many other languages would likely require a re-assessment of our use of orthography as a proxy for linguistic form. The assumption that orthography is a reasonable proxy for form is not accurate in many languages – however, at present UniMorph does not include phonological transcriptions, and automated grapheme-to-phoneme conversion across a broad range of languages is the subject of very active research (Ashby *et al.* 2021). These difficulties would need to be overcome in order to use phonological transcriptions. Future work should also investigate to what degree our variability of embedding measure is equivalent to or complementary to Copot *et al.*'s (2022) predictability of frequency measure, as both are motivated from semantic drift due to a change in lexical index. Similarly, future work could clarify the contribution of distributional semantics by using a model such as Word2Vec or GloVe, or newer models of distributional semantics, such as XLM-R (Conneau *et al.* 2020) – though in the latter case they would have to overcome the difficulties of multilingual decontextualisation as described in Section 3.2. Further, as we use only two simple classification techniques (logistic regression and an MLP), it is possible that further hyperparameter tuning or use of other techniques, such as random forests or gradient boosting, could improve on classification accuracy.

In this work, we have presented the first multilingual computational study of the inflection–derivation distinction. In Section 3 we define a small set of measures capturing the hypothesised tendency of derivation to produce bigger and more variable changes to the base form in terms of form, syntax, and semantics. We then systematically study the relationship between these measures and traditional categorisations of

morphological constructions into inflection and derivation, which we derive from the UniMorph 4.0 dataset. In Section 5, we show that these measures each correlate, in some cases strongly, with whether a construction is listed as inflectional or derivational in UniMorph 4.0. We show evidence that these correlations are not due to systematic differences in the frequency of inflectional and derivational constructions. In Section 6, we show that both logistic regression and multi-layer perceptron classifiers which use these measures as inputs can be trained to reconstruct most of the UniMorph inflection–derivation distinction, with logistic classifier achieving a classification accuracy of $83 \pm 1\%$ and the MLP achieving a classification accuracy of $89 \pm 1\%$, improving by 26 and 32 points over predicting the majority class, respectively. We identify the variability of the change in distributional embedding space V_{Embed} and the variability of the change of form V_{Form} as particularly strong correlates of the distinction, together able to classify $83 \pm 1\%$ of constructions as they are classified in UniMorph.

Overall, these results show that much of the categories of inflection and derivation as used in UniMorph can be accounted for by corpus-based measures which make concrete the subjective tests suggested by linguists. In so doing, we have also validated in a larger, multilingual context the core findings of Bonami and Paperno (2018) and Rosa and Žabokrtský (2019), finding that these properties hold across 26 languages (21 Indo-European and 5 others), with a model that uses no language-specific features. These well-defined, empirical measures avoid the often-discussed subjectivity and vagueness of existing criteria (Haspelmath 2024; Plank 1994; Bybee 1985), and enable us to produce the first large-scale quantification of how consistently the categories of inflection and derivation are applied, and validate that these measures can *generalise* to unseen constructions.

With these measures, we are also able to identify in a quantitative way *how canonical* different categories of inflections are (Section 7) in terms of properties of their form and distribution. We determine, that, as suggested by Booij (1996), inherent inflection is a *non-canonical inflectional category* under our model: inflectional constructions which are purely inherent are significantly more likely to be classified as derivations than other inflections under our model. We find in our sample this seems to be particularly due to *nominal* inherent inflections, like case and number. We find no traditional categories of

inflectional meaning significantly non-canonical, providing some validation accounts of inflection which are structured around these categories like Haspelmath (2024) or Sylak-Glassman (2016), though we find weak evidence that voice and comparatives could be such categories.

Finally, we note that while there is a high degree of consistency in the use of the terms inflection and derivation in terms of our measures and combining multiple measures reduces the amount of overlap between inflectional and derivational constructions, we still find many constructions near the model’s decision boundary between the two categories, indicating a gradient, rather than categorical, distinction (Section 8.4). This gradient region is relatively small, as suggested by our high accuracies, but does not suggest inflection and derivation as categories *naturally emerging* from our measures.

ACKNOWLEDGEMENTS

The authors would like to thank Paul J.W. Schauenburg, Albert Haley, Itamar Kastner, Kate McCurdy, and Francis Mollica for their comments on this work. This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk). This work was in part supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

REFERENCES

Peter ACKEMA and Ad NEELEMAN (2019), *Default person versus default number in agreement*, pp. 21–54, Open Generative Syntax, Language Science Press, doi:10.5281/zenodo.3458062.

Stephen R. ANDERSON (1982), Where's morphology?, *Linguistic Inquiry*, 13:571–612.

Stephen R. ANDERSON (1985), Inflectional morphology, in *Language Typology and Syntactic Description*, volume 3, pp. 150–201, Cambridge University Press, 1 edition.

Antti ARPPE, Atticus HARRIGAN, Katherine SCHMIRLER, Lene ANTONSEN, Trond TROSTERUD, Sjur NØRSTEBØ MOSHAGEN, Miikka SILFVERBERG, Arok WOLVENGREY, Conor SNOEK, Jordan LACHLER, Eddie Antonio SANTOS, Jean OKIMĀSIS, and Dorothy THUNDER (2014–2019), Finite-state transducer-based computational model of Plains Cree morphology, <https://giellalt.uit.no/lang/crk/PlainsCreeDocumentation.html>.

Lucas F.E. ASHBY, Travis M. BARTLEY, Simon CLEMATIDE, Luca DEL SIGNORE, Cameron GIBSON, Kyle GORMAN, Yeonju LEE-SIKKA, Peter MAKAROV, Aidan MALANOSKI, Sean MILLER, Omar ORTIZ, Reuben RAFF, Arundhati SENGUPTA, Bora SEO, Yulia SPEKTOR, and Winnie YAN (2021), Results of the second SIGMORPHON shared task on multilingual grapheme-to-phoneme conversion, in *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 115–125, Association for Computational Linguistics, Online, doi:10.18653/v1/2021.sigmorphon-1.13, <https://aclanthology.org/2021.sigmorphon-1.13>.

Madina BABAZHANOVA, Maxat TEZEKBAYEV, and Zhenisbek ASSYLBEKOV (2021), Geometric probing of word vectors, in *ESANN 2021 Proceedings – 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 587–592, i6doc.com publication, Virtual, Online, Belgium, doi:10.14428/esann/2021.ES2021-105.

Khuyagbaatar BATSUREN, Gábor BELLA, and Fausto GIUNCHIGLIA (2021), MorphyNet: a large multilingual database of derivational and inflectional morphology, in *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 39–48, Association for Computational Linguistics, Online, doi:10.18653/v1/2021.sigmorphon-1.5, <https://aclanthology.org/2021.sigmorphon-1.5>.

Khuyagbaatar BATSUREN, Omer GOLDMAN, Salam KHALIFA, Nizar HABASH, Witold KIERAŚ, Gábor BELLA, Brian LEONARD, Garrett NICOLAI, Kyle GORMAN, Yustinus Ghanggo ATE, Maria RYSKINA, Sabrina MIELKE, Elena BUDIANSKAYA, Charbel EL-KHAISSI, Tiago PIMENTEL, Michael GASSER, William Abbott LANE, Mohit RAJ, Matt COLER, Jaime Rafael Montoya SAMAME, Delio Siticonatzi CAMAITERI, Esaú Zumaeta ROJAS, Didier LÓPEZ FRANCIS, Arturo ONCEVAY, Juan LÓPEZ BAUTISTA, Gema Celeste Silva VILLEGAS, Lucas Torroba HENNIGEN, Adam EK, David GURIEL, Peter DIRIX, Jean-Philippe BERNARDY, Andrey SCHERBAKOV, Aziyana BAYYR-OOL, Antonios ANASTASOPOULOS, Roberto ZARIQUIEY, Karina SHEIFER, Sofya

GANIEVA, Hilaria CRUZ, Ritván KARAHÓGA, Stella MARKANTONATOU, George PAVLIDIS, Matvey PLUGARYOV, Elena KLYACHKO, Ali SALEHI, Candy ANGULO, Jatayu BAXI, Andrew KRIZHANOVSKY, Natalia KRIZHANOVSKAYA, Elizabeth SALESKY, Clara VANIA, Sardana IVANOVA, Jennifer WHITE, Rowan Hall MAUDSLAY, Josef VALVODA, Ran ZMIGROD, Paula CZARNOWSKA, Irene NIKKARINEN, Aelita SALCHAK, Brijesh BHATT, Christopher STRAUGHN, Zoey LIU, Jonathan North WASHINGTON, Yuval PINTER, Duygu ATAMAN, Marcin WOLINSKI, Totok SUHARDIJANTO, Anna YABLONSKAYA, Niklas STOEHR, Hossep DOLATIAN, Zahroh NURIAH, Shyam RATAN, Francis M. TYERS, Edoardo M. PONTI, Grant AITON, Aryaman ARORA, Richard J. HATCHER, Ritesh KUMAR, Jeremiah YOUNG, Daria RODIONOVA, Anastasia YEMELINA, Taras ANDRUSHKO, Igor MARCHENKO, Polina MASHKOVTSOVA, Alexandra SEROVA, Emily PRUD'HOMMEAUX, Maria NEPOMNIASHCHAYA, Fausto GIUNCHIGLIA, Eleanor CHODROFF, Mans HULDEN, Miikka SILFVERBERG, Arya D. MCCARTHY, David YAROWSKY, Ryan COTTERELL, Reut TSARFATY, and Ekaterina VYLOMOVA (2022), UniMorph 4.0: Universal Morphology, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 840–855, European Language Resources Association, Marseille, France, <https://aclanthology.org/2022.lrec-1.89>.

Laurie BAUER (2004), The function of word-formation and the inflection-derivation distinction, *Words and their Places. A Festschrift for J. Lachlan Mackenzie*. Amsterdam: Vrije Universiteit, pp. 283–292.

Sacha BENIAMINE, Martin MAIDEN, and Erich ROUND (2020), Opening the Romance verbal inflection dataset 2.0: A CLDF lexicon, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3027–3035, European Language Resources Association, Marseille, France, <https://aclanthology.org/2020.lrec-1.370>.

Toms BERGMANIS and Sharon GOLDWATER (2017), From segmentation to analyses: a probabilistic model for unsupervised morphology induction, in *Proceedings of EACL*, Valencia, Spain.

Piotr BOJANOWSKI, Edouard GRAVE, Armand JOULIN, and Tomas MIKOLOV (2017), Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics*, 5:135–146, doi:10.1162/tacl_a_00051, <https://aclanthology.org/Q17-1010>.

Rishi BOMMASANI, Kelly DAVIS, and Claire CARDIE (2020), Interpreting pretrained contextualized representations via reductions to static embeddings, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4758–4781, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.acl-main.431, <https://aclanthology.org/2020.acl-main.431>.

Olivier BONAMI and Denis PAPERNO (2018), Inflection vs. derivation in a distributional vector space, *Lingue e linguaggio*, 17(2):173–196.

Olivier BONAMI and Jana STRNADOVÁ (2019), Paradigm structure and predictability in derivational morphology, *Morphology*, 29(2):167–197, ISSN 1871-5656, doi:10.1007/s11525-018-9322-6.

Geert BOOIJ (1996), Inherent versus contextual inflection and the split morphology hypothesis, in *Yearbook of Morphology 1995*, pp. 1–16, Springer.

Geert BOOIJ (2007), Inflection, in *The Grammar of Words: An Introduction to Linguistic Morphology*, Oxford University Press, doi:10.1093/acprof:oso/9780199226245.003.0005.

R. D. BOSCHLOO (1970), Raised conditional level of significance for the 2×2-table when testing the equality of two probabilities, *Statistica Neerlandica*, 24(1):1–9.

Joan L. BYBEE (1985), *Morphology: A study of the relation between meaning and form*, John Benjamins, Amsterdam.

Alexis CONNEAU, Kartikay KHANDELWAL, Naman GOYAL, Vishrav CHAUDHARY, Guillaume WENZKE, Francisco GUZMÁN, Edouard GRAVE, Myle OTT, Luke ZETTMAYER, and Veselin STOYANOV (2020), Unsupervised cross-lingual representation learning at scale, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.acl-main.747, <https://aclanthology.org/2020.acl-main.747>.

Maria COPOT, Timothee MICKUS, and Olivier BONAMI (2022), Idiosyncratic frequency as a measure of derivation vs. inflection, *Journal of Language Modelling*, 10(2):193–240, doi:10.15398/jlm.v10i2.301, <https://jlm.ipipan.waw.pl/index.php/JLM/article/view/301>.

Greville G. CORBETT (2010), Canonical derivational morphology, *Word Structure*, 3(2):141–155.

Ryan COTTERELL and Hinrich SCHÜTZE (2018), Joint semantic synthesis and morphological analysis of the derived word, *Transactions of the Association for Computational Linguistics*, 6:33–48, doi:10.1162/tacl_a_00003, <https://aclanthology.org/Q18-1003>.

William CROFT (2002), *Typology and universals*, Cambridge Textbooks in Linguistics, Cambridge University Press, 2 edition, doi:10.1017/CBO9780511840579.

Anne CUTLER (1981), Degrees of transparency in word formation, *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 26(1):73–77.

Daniel DEUTSCH, John HEWITT, and Dan ROTH (2018), A distributional and orthographic aggregation model for English derivational morphology, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1938–1947, Association for

Computational Linguistics, Melbourne, Australia, doi:10.18653/v1/P18-1180, <https://aclanthology.org/P18-1180>.

Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE, and Kristina TOUTANOVA (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota, doi:10.18653/v1/N19-1423, <https://aclanthology.org/N19-1423>.

Wolfgang U. DRESSLER (1989), Prototypical differences between inflection and derivation, *STUF – Language Typology and Universals*, 42(1):3–10.

Matthew S. DRYER (1989), Large linguistic areas and language sampling, *Studies in Language. International Journal sponsored by the Foundation “Foundations of Language”*, 13(2):257–292.

Edouard GRAVE, Piotr BOJANOWSKI, Prakhar GUPTA, Armand JOULIN, and Tomas MIKOLOV (2018), Learning word vectors for 157 languages, in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan, <https://aclanthology.org/L18-1550>.

Joseph H. GREENBERG, editor (1966), *Universals of language*, M.I.T. Press, 2 edition.

P. HACKEN (1994), *Defining morphology: A principled approach to determining the boundaries of compounding, derivation, and inflection*, Altermumswissenschaftliche Texte Und Studien, G. Olms Verlag, https://books.google.co.uk/books?id=E8mWh_6mRACc.

Zellig HARRIS (1954), Distributional structure, *Word*, 10(23):146–162.

Martin HASPELMATH (2024), Inflection and derivation as traditional comparative concepts, *Linguistics*, 62(1):43–77, doi:doi:10.1515/ling-2022-0086, <https://doi.org/10.1515/ling-2022-0086>.

Nabil HATHOUT and Fiammetta NAMER (2016), Giving lexical resources a second life: Démonette, a multi-sourced morpho-semantic network for French, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pp. 1084–1091, European Language Resources Association (ELRA), Portorož, Slovenia, <https://aclanthology.org/L16-1173>.

Nabil HATHOUT, Franck SAJOUS, and Basilio CALDERONE (2014), GLÀFF, a large versatile French lexicon, in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pp. 1007–1012, European Language Resources Association (ELRA), Reykjavik, Iceland, http://www.lrec-conf.org/proceedings/lrec2014/pdf/58_Paper.pdf.

- Junxian HE, Graham NEUBIG, and Taylor BERG-KIRKPATRICK (2018), Unsupervised learning of syntactic structure with invertible neural projections, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1292–1302, Association for Computational Linguistics, Brussels, Belgium, doi:10.18653/v1/D18-1160, <https://aclanthology.org/D18-1160>.
- Valentin HOFMANN, Hinrich SCHÜTZE, and Janet PIERREHUMBERT (2020), A graph auto-encoder model of derivational morphology, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1127–1138, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.acl-main.106, <https://aclanthology.org/2020.acl-main.106>.
- Zongliang HU, Kai DONG, Wenlin DAI, and Tiejun TONG (2017), A comparison of methods for estimating the determinant of high-dimensional covariance matrix, *The International Journal of Biostatistics*, 13(2):20170013, doi:10.1515/ijb-2017-0013.
- M. KASTHURI, S. Britto Ramesh KUMAR, and Souheil KHADDAJ (2017), PLIS: Proposed language independent stemmer for information retrieval systems using dynamic programming, in *2017 World Congress on Computing and Communication Technologies (WCCCT)*, pp. 132–135, doi:10.1109/WCCCT.2016.39.
- Diederik P. KINGMA and Jimmy BA (2015), Adam: A method for stochastic optimization, in Yoshua BENGIO and Yann LECUN, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*, <http://arxiv.org/abs/1412.6980>.
- Bilal KIRKICI and Harald CLAHSSEN (2013), Inflection and derivation in native and non-native language processing: Masked priming experiments on Turkish, *Bilingualism: Language and Cognition*, 16(4):776–791, doi:10.1017/S1366728912000648.
- Christa KÖNIG (2006), Marked nominative in Africa, *Studies in Language. International Journal sponsored by the Foundation “Foundations of Language”*, 30(4):655–732.
- Lukáš KYJÁNEK, Zdeněk ŽABOKRTSKÝ, Magda ŠEVČÍKOVÁ, and Jonáš VIDRA (2020), Universal Derivations 1.0, a growing collection of harmonised word-formation resources, *The Prague Bulletin of Mathematical Linguistics*, 2(115):333–348.
- Lior LAKS and Fiammetta NAMER (2022), Hebrewnette – a new derivational resource for non-concatenative morphology: Principles, design and implementation, *The Prague Bulletin of Mathematical Linguistics*, 118:25–53.
- Septina Dian LARASATI, Vladislav KUBOŇ, and Daniel ZEMAN (2011), Indonesian morphology tool (MorphInd): Towards an Indonesian corpus, in

- Cerstin MAHLOW and Michael PIOTROWSKI, editors, *Systems and Frameworks for Computational Morphology*, pp. 119–129, Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-642-23138-4_8.
- Alessandro LAUDANNA, William BADECKER, and Alfonso CARAMAZZA (1992), Processing inflectional and derivational morphology, *Journal of Memory and Language*, 31(3):333–348.
- Vladimir LEVENSHTAIN (1966), Binary codes capable of correcting deletions, insertions and reversals, *Soviet Physics Doklady*, 10:707.
- Chu-Cheng LIN, Waleed AMMAR, Chris DYER, and Lori LEVIN (2015), Unsupervised POS induction with word embeddings, in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1311–1316, Association for Computational Linguistics, Denver, Colorado, doi:10.3115/v1/N15-1144, <https://aclanthology.org/N15-1144>.
- Nikola LJUBEŠIĆ, Filip KLUBIČKA, Željko AGIĆ, and Ivo-Pavao JAZBEC (2016), New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4264–4270, European Language Resources Association (ELRA), Portorož, Slovenia, <https://aclanthology.org/L16-1676>.
- Donald G. MACKAY (1978), Derivational rules and the internal lexicon, *Journal of Verbal Learning and Verbal Behavior*, 17(1):61–71.
- Robert MALOUF, Farrell ACKERMAN, and Arturs SEMENUKS (2020), Lexical databases for computational analyses: A linguistic perspective, in Allyson ETTINGER, Gaja JAROSZ, and Joe PATER, editors, *Proceedings of the Society for Computation in Linguistics 2020*, pp. 446–456, Association for Computational Linguistics, New York, New York, <https://aclanthology.org/2020.scil-1.52>.
- Arya D. MCCARTHY, Christo KIROV, Matteo GRELLA, Amrit NIDHI, Patrick XIA, Kyle GORMAN, Ekaterina VYLOMOVA, Sabrina J. MIELKE, Garrett NICOLAI, Miikka SILFVERBERG, Timofey ARKHANGELSKIY, Nataly KRIZHANOVSKY, Andrew KRIZHANOVSKY, Elena KLYACHKO, Alexey SOROKIN, John MANSFIELD, Valts ERNŠTREITS, Yuval PINTER, Cassandra L. JACOBS, Ryan COTTERELL, Mans HULDEN, and David YAROWSKY (2020), UniMorph 3.0: Universal Morphology, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3922–3931, European Language Resources Association, Marseille, France, <https://aclanthology.org/2020.lrec-1.483>.
- Tomas MIKOLOV, Ilya SUTSKEVER, Kai CHEN, Greg CORRADO, and Jeffrey DEAN (2013), Distributed representations of words and phrases and their compositionality, in *Proceedings of the 26th International Conference on Neural*

Information Processing Systems – Volume 2, NIPS’13, p. 3111–3119, Curran Associates Inc., Red Hook, NY, USA.

Karthik NARASIMHAN, Regina BARZILAY, and Tommi JAAKKOLA (2015), An unsupervised method for uncovering morphological chains, *Transactions of the Association for Computational Linguistics*, 3:157–167.

Bruce OLIVER, Clarissa FORBES, Changbing YANG, Farhan SAMIR, Edith COATES, Garrett NICOLAI, and Miikka SILFVERBERG (2022), An inflectional database for Gitksan, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6597–6606, European Language Resources Association, Marseille, France, <https://aclanthology.org/2022.lrec-1.710>.

Jeffrey PENNINGTON, Richard SOCHER, and Christopher MANNING (2014), GloVe: Global vectors for word representation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Association for Computational Linguistics, Doha, Qatar, doi:10.3115/v1/D14-1162, <https://aclanthology.org/D14-1162>.

David PERLMUTTER (1988), The split morphology hypothesis: Evidence from Yiddish, *Theoretical Morphology*, pp. 79–100.

Tiago PIMENTEL, Josef VALVODA, Rowan Hall MAUDSLAY, Ran ZMIGROD, Adina WILLIAMS, and Ryan COTTERELL (2020), Information-theoretic probing for linguistic structure, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.acl-main.420, <https://aclanthology.org/2020.acl-main.420>.

Frans PLANK (1994), Inflection and derivation, in *The Encyclopedia of Language and Linguistics*, pp. 1671–1679, Elsevier Science and Technology, Amsterdam.

Shauli RAVFOGEL, Yanai ELAZAR, Jacob GOLDBERGER, and Yoav GOLDBERG (2020), Unsupervised distillation of syntactic information from contextualized word representations, in *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 91–106, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.blackboxnlp-1.9, <https://aclanthology.org/2020.blackboxnlp-1.9>.

Rudolf ROSA and Zdeněk ŽABOKRTSKÝ (2019), Attempting to separate inflection and derivation using vector space representations, in *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pp. 61–70, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czechia, <https://aclanthology.org/W19-8508>.

Rudolf ROSA and Zdenek ZABOKRTSKÝ (2019), Unsupervised lemmatization as embeddings-based word clustering, *CoRR*, abs/1908.08528, <http://arxiv.org/abs/1908.08528>.

- Jenny R. SAFFRAN, Richard N. ASLIN, and Elissa L. NEWPORT (1996), Statistical learning by 8-month-old infants, *Science*, 274(5294):1926–1928.
- Adriaan M. J. SCHAKEL and Benjamin J. WILSON (2015), Measuring word significance using distributed representations of words, *Computing Research Repository*, arXiv:1508.02297, <http://arxiv.org/abs/1508.02297>.
- Patrick SCHONE and Daniel JURAFSKY (2000), Knowledge-free induction of morphology using latent semantic analysis, in *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*, <https://aclanthology.org/W00-0712>.
- Michael SILVERSTEIN (1986), Hierarchy of features and ergativity, in *Features and Projections*, pp. 163–232, De Gruyter Mouton, Berlin, Boston, doi:10.1515/9783110871661-008.
- Radu SORICUT and Franz OCH (2015), Unsupervised morphology induction using word embeddings, in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1627–1637, Association for Computational Linguistics, Denver, Colorado, doi:10.3115/v1/N15-1186, <https://aclanthology.org/N15-1186>.
- Andrew SPENCER (2013), *Lexical relatedness*, Oxford University Press, Oxford.
- Pavol ŠTEKAUER (2015), 14. The delimitation of derivation and inflection, in Peter O. MÜLLER, Ingeborg OHNHEISER, Susan OLSEN, and Franz RAINER, editors, *Volume 1 Word-Formation*, pp. 218–235, De Gruyter Mouton.
- Lonny Alaskuk STRUNK (2020), *A finite-state morphological analyzer for Central Alaskan Yup'ik*, University of Washington.
- Daniel SWINGLEY (2005), Statistical clustering and the contents of the infant vocabulary, *Cognitive Psychology*, 50(1):86–132.
- John SYLAK-GLASSMAN (2016), The composition and use of the universal morphological feature schema (UniMorph schema), <https://unimorph.github.io/doc/unimorph-schema.pdf>.
- Erik D. THIESSEN, Alexandra T. KRONSTEIN, and Daniel G. HUFNAGLE (2013), The extraction and integration framework: a two-process account of statistical learning, *Psychological Bulletin*, 139(4):792.
- Erik D. THIESSEN and Jenny R. SAFFRAN (2003), When cues collide: use of stress and statistical cues to word boundaries by 7-to-9-month-old infants, *Developmental Psychology*, 39(4):706.
- Susan P. THOMPSON and Elissa L. NEWPORT (2007), Statistical learning of syntax: The role of transitional probability, *Language Learning and Development*, 3(1):1–42.

Hugo David Calderon VILCA, Flor Cagniy Cárdenas MARINÓ, and Edwin Fredy Mamani CALDERON (2012), *Analizador morfológico de la lengua Quechua basado en software libre Helsinki-finite-state-transducer (HFST)*.

Ivan VULIĆ, Simon BAKER, Edoardo Maria PONTI, Ulla PETTI, Ira LEVIANT, Kelly WING, Olga MAJEWSKA, Eden BAR, Matt MALONE, Thierry POIBEAU, Roi REICHART, and Anna KORHONEN (2020), Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity, *Computational Linguistics*, 46(4):847–897, doi:10.1162/coli_a_00391, <https://aclanthology.org/2020.cl-4.5>.

Christian WARTENA (2013), Distributional similarity of words with different frequencies, in *Proceedings of the 13th edition of the Dutch-Belgian information retrieval Workshop (DIR 2013)*, pp. 8–11, Hochschule Hannover.

Adam WIEMERSLAGE, Arya D. MCCARTHY, Alexander ERDMANN, Garrett NICOLAI, Manex AGIRREZABAL, Miikka SILFVERBERG, Mans HULDEN, and Katharina KANN (2021), Findings of the SIGMORPHON 2021 shared task on unsupervised morphological paradigm clustering, in *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 72–81.

Coleman Haley

© 0000-0003-3089-9558
coleman.haley@ed.ac.uk

Sharon Goldwater

© 0000-0002-7298-0947
sgwater@inf.ed.ac.uk

Edoardo M. Ponti

© 0000-0002-6308-1050
eponti@ed.ac.uk

Institute for Language, Cognition
and Computation
School of Informatics
University of Edinburgh
Edinburgh, UK

Coleman Haley, Edoardo M. Ponti, and Sharon Goldwater (2024), *Corpus-based measures discriminate inflection and derivation cross-linguistically*, *Journal of Language Modelling*, 12(2):477–529

doi <https://dx.doi.org/10.15398/jlm.v12i2.351>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

cc  <http://creativecommons.org/licenses/by/4.0/>