

Detecting Lexical Borrowings from Dominant Languages in Multilingual Wordlists

John E. Miller

Artificial Intelligence/Engr. Chair of Multilingual Computational Linguistics
Pontificia Universidad Católica del Perú
San Miguel, Lima, Peru
jemiller@pucp.edu.pe

Johann-Mattis List

University of Passau
Germany
mattis.list@uni-passau.de

Abstract

Language contact is a pervasive phenomenon reflected in the borrowing of words from donor to recipient languages. Most computational approaches to borrowing detection treat all languages under study as equally important, even though dominant languages have a stronger impact on heritage languages than vice versa. We test new methods for lexical borrowing detection in contact situations where dominant languages play an important role, applying two classical sequence comparison methods and one machine learning method to a sample of seven Latin American languages which have all borrowed extensively from Spanish. All methods perform well, with the supervised machine learning system outperforming the classical systems. A review of detection errors shows that borrowing detection could be substantially improved by taking into account donor words with divergent meanings from recipient words.

1 Introduction

Language contact is one of the most pervasive linguistic phenomena. It is the first factor that needs to be excluded when searching for genetic relations among the languages of the world, and it needs to be controlled for when searching for cross-linguistic universals. Any study trying to explain how humans use language must take language contact into account. It is also a unique witness of ancient contacts in human prehistory.

Language contact is most prominently reflected in *lexical borrowing*, the *transfer* of words from a donor language to a recipient language. Although research on the computational handling of lexical borrowing has made some progress of late, computational approaches to the investigation of language contact are still in their infancy (List, 2019). Specifically methods that could infer the direction of borrowings have not been proposed so far. While numerous case studies investigate the influence of

dominant languages on heritage languages (Prochazka and Vogl, 2017; Meisel, 2018), we lack automated methods that could be used to study the influence of particular dominant languages in linguistically diverse areas of the world.

The number of standardized cross-linguistic wordlist collections has greatly increased over the past years (Rzyski et al., 2020; List et al., 2022a), with standard formats (Forkel et al., 2018) accompanied by software tools that allow scholars to prepare and curate standardized wordlists in an efficient manner (Forkel and List, 2020). It would be useful to have a computer assisted tool for linguists to detect which words in a cross-linguistic region have been borrowed from a dominant language and with further advances, a tool for inclusion in language contact assessment or in computational cladistics workflows.

Here, we compare the suitability of three different systems to address this task – two systems based on classical algorithms for automated sequence comparison that can be applied in supervised and unsupervised settings, and one supervised system based on extended machine learning techniques. We test these systems on a newly derived dataset of seven Latin American languages (Fig. 1) in which Spanish is a dominant donor language. Our results show that the supervised machine-learning system outperforms the classical systems.

2 Previous Work

Although still few in number, automatic methods for borrowing detection have been increasingly applied and developed in the past years. Early studies by van der Ark et al. (2007) and later Menecier et al. (2016) compute edit distances between words from genetically unrelated languages and compare distances to thresholds, in order to detect borrowed words in multilingual wordlists.

Besides edit distance, which directly calculates distances between phonetic sequences, sound class

based methods cluster phonetic segments into sound classes and then compute distances between sound class sequences. The sound-class based alignment method (SCA) (List, 2012) provides sound class categories, scoring functions for distance measures, and modifiable gap scores based on prosodic context.

Zhang et al. (2021) compare edit distance performance against SCA (List, 2012) distance performance, finding that SCA outperforms edit distance in accuracy. Hantgan et al. (2022) build on this work, using dedicated methods for automated cognate detection applied to languages from different language families in order to identify clusters of related words resulting from lexical borrowing. List and Forkel (2022a) expand this work further, by applying a two-stage workflow in which they first identify language-family-internal cognates, using a method specifically apt for the detection of deep cognates, and then compute SCA distances between cognate sets from genetically unrelated languages in order to infer sets of words related by lexical transfer.

Miller et al. (2020) compute language models for inherited and borrowed words for individual languages from the World Loanword Database (WOLD, Haspelmath and Tadmor 2009) using Markov Chains and Recursive Neural Networks and compare cross-entropies for inherited and borrowed language models in order to identify borrowings from monolingual information alone.

Kaipng and Klamer (2022) use automated methods for cognate detection (List et al., 2017) on a target set of Timor-Alor-Pantar languages. In order to infer borrowings from Indonesian and Tetun (not in the target set), they include both languages in their sample and treat all cognate sets that involves words from either of the two languages as borrowings. Moro et al. (2023) apply a similar approach to investigate borrowings in Alorese.

Mi et al. (2020) and Nath et al. (2022) train binary classifiers, mainly neural based, on large wordlists to predict borrowed words, and achieve F1 scores in the 0.75 to 0.85 range. Their workflows seem cumbersome, compute intensive, and not minimalist, but the results are promising.

3 Materials and Methods

3.1 Materials

For this study, a new comparative wordlist was created by taking data for seven Latin American lan-

guages from WOLD (<https://wold.clld.org>, Haspelmath and Tadmor 2009) and combining them with a wordlist of Spanish derived from the Intercontinental Dictionary Series (<https://ids.clld.org>, Key and Comrie 2015). Phonetic transcriptions for the Latin American languages were added to WOLD by Miller et al. (2020). Latin American Spanish phonetic transcriptions were added for this study (and could be later expanded by adding more transcriptions from historical varieties of Spanish). The resulting dataset conforms to the standards suggested by the Cross-Linguistic Data Formats initiative (CLDF, <https://clfd.clld.org>, Forkel et al. 2018). The data curation follows the Lexibank workflow (List et al., 2022a) and checks that data conform to certain standards, with languages being linked to Glottolog (<https://glottolog.org>, Hammarström et al. 2022, Version 4.7), concepts being linked to Concepticon (<https://concepticon.clld.org>, List et al. 2022b, Version 3.0), and transcriptions following the B(road)IPA conventions of the Cross-Linguistic Transcription Systems reference catalog (<https://clts.clld.org>, List et al. 2021, Version 2.2, see Anderson et al. 2018). Details of the resulting database are shown in the map of language locations along with percentages for borrowings from Spanish in Fig. 1 and in Tab. 1. Q’eqchi’ and Zinacantan Tzotzil are both Mayan languages, but appear substantially varied in the database.

Language	Concepts	Lexemes	Segments	Vocab.
Imb. Quechua	1,155	1,156	7,177	33
Mapudungun	1,040	1,242	7,356	33
Otomi	1,252	2,241	11,730	57
Q’eqchi’	1,211	1,773	10,367	49
Wichí	1,128	1,219	8,233	44
Yaqui	1,242	1,433	9,297	28
Zin. Tzotzil	955	1,266	7,129	41
Spanish	1,308	1,770	11,261	30
Aggregate	1,308	12,100	72,550	112

Table 1: Database details for seven Latin American languages plus Latin American Spanish.

3.2 Methods

Methods for Borrowing Detection. We develop three different methods for the detection of borrowings *from* a dominant language *to* non-dominant languages in multilingual wordlists. Following historical linguistics comparative method practice (Campbell, 2013), only word forms corresponding to the same concept are considered as

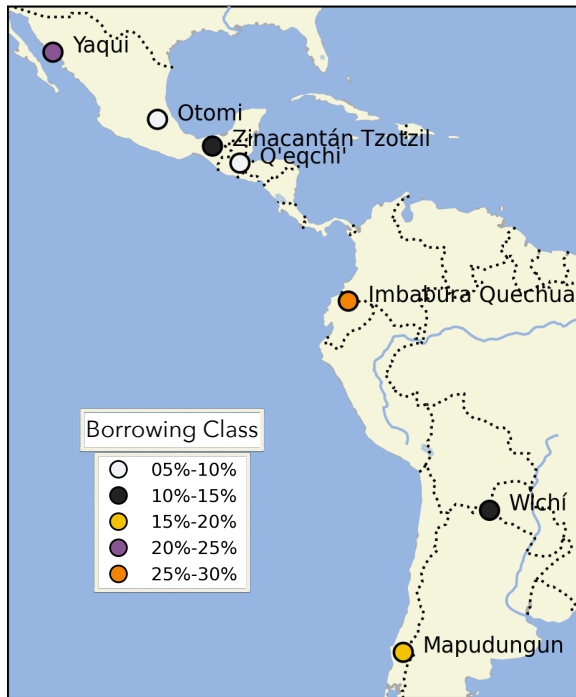


Figure 1: Map of languages with Spanish borrowing class.

candidates for borrowing.

The first method, called *Closest Match* borrowing detection in the following, iterates over all word pairs that express the same concept in the dominant language and the heritage languages and then computes phonetic distances. Word pairs whose phonetic distance is below a certain threshold are judged to be borrowings from the dominant language. We test two phonetic distances, the normalized edit distance (NED) – the classical edit distance (Levenshtein, 1965) between two words, divided by the length of the longer word – and the SCA distance (List, 2012).

The second method, called *Cognate-Based* borrowing detection in the following, follows the approach by Hantgan et al. (2022): it first computes cognates using a cluster-based approach for automated cognate detection in which words expressing the same concept whose average phonetic distance is below a certain threshold are assigned to the same cognate set (List et al., 2017), and then identifies all words assigned to cognate sets involving the dominant language as borrowings. We tested again normalized edit and SCA distances.

The third method, called *Classifier-Based* borrowing detection in the following, iterates over all word pairs with the same concept, but stores phonetic distance scores for various distance measures as vectors, which can then be used to train a classi-

fier, firstly, a linear Support Vector Machine (SVM) (Cristianini and Shawe-Taylor, 2000), in a supervised setting. We tested various phonetic distance measures, but report only on the combination of normalized edit and SCA distances, as they yielded the best results. Both Closest Match and Cognate-Based methods require a fixed threshold which we estimate from the training data. So all three methods are considered as supervised.

Sampling. Train-test splits are made based on concepts rather than individual word entries. This permits matching of words for the same concept in all methods without loss of candidate words. Treating train-test split as a nuisance variable takes into account differences between partitions across methods thus controlling for effects of sampling by concepts with differing borrowing behavior or statistical dependencies between test partitions due to sampling without replacement. See (Dror et al., 2018) for mention of the dependency problem with cross-validation. Our use of a fixed partition across treatments and analysis of variance controlling for partition as a nuisance or ‘blocking’ variable accounts for this dependency, and takes advantage of any systematic effects in borrowing behavior by partition.

Evaluation. For the analysis of the cross-validation data, we use a randomized blocks design where experiment is the treatment or factor, and test partition is the randomized block or nuisance variable. A standard analysis of variance partitions treatment effects, nuisance variable, and error, and permits a more powerful test of treatment differences without the nuisance variable variance. We follow up statistically significant findings for treatment (experiment), with comparisons of experiments versus the overall average using a joint (family) error rate.

Implementation. Our methods are implemented in Python, making specifically use of the CLDF-Bench package (<https://pypi.org/project/cldfbench/>, Forkel and List 2020, Version 1.13.0) to provide commandline access to all methods described here. For the computation of alignments and edit distances, LingPy (<https://pypi.org/project/lingpy>, List and Forkel 2022b, Version 2.6.9) is used. SVM and evaluation are realized with the help of Scikit-Learn (<https://pypi.org/project/scikit-learn/>, Pedregosa et al. 2011, Version 1.2.1).

4 Results and Discussion

We tested our three methods with two distance measures in five experiments (normalized edit and SCA distances individually in both Closest Match and Cognate-Based methods, and combined in the Classifier-Based method) using a 10-fold cross-validation on our data and reporting precision, recall, F1 scores, accuracy, and execution times (mm:ss). F1 score is the primary result measure; accuracy and execution time are informational. The 10-fold cross-validation uses the same 10 fixed train and non-overlapping test splits for all experiments. With few parameter estimates (1 threshold each for Closest Match and Cognate-Based, 2 distance and 7 target language coefficients for Classifier), a separate *train* split into *fit/val* is not necessary.

All methods perform well with less than 5 points separating the highest from the lowest F1 scores. Tab. 2 shows the results of the ten-fold cross validation of our three methods in five experiments. An analysis of variance,¹ with experiment as the effects variable and train-test split as the nuisance variable, shows highly significant effects for precision ($F_{4,36} = 25.74, p < 0.0001$) and F1 score ($F_{4,36} = 14.3, p < 0.0001$).

Closest Match with normalized edit distance performs poorly, while Classifier-Based with combined normalized edit and SCA distances performs well. Classifier-Based performs better than the average of all experiments in F1 score, and substantially better in precision versus other experiments; the method is conservative, with a low number of false positives. Performance on remaining experiments is indistinguishable from the overall average of all experiments combined. The Cognate-Based method is compute intensive performing multiple alignment over all languages. Accuracy is well above the majority decision accuracy of 84.8% (100% – 15.2% borrowing) in all experiments.

A search for classifier improvements prompted several *ad hoc* experiments (see tab. 3). We observe: (1) A radial basis function (rbf) SVM classifier performs no better than our linear SVM. We suspect the estimated target language parameters do not generalize well to held-out data. (2) A logistic regression classifier performs on par with our linear SVM. (3) A weight balanced SVM classifier trades an increase in recall for a larger drop in precision. We also test whether using separate trials for each target language in Closest Match,

¹Statistical analyses with JMP (SAS Institute Inc., 2021).

Method	Prec.	Rec.	F1	Acc.	mm:ss
Closest Match					
NED	<u>0.832</u>	0.703	<u>0.761</u>	0.938	00:15
SCA	0.869	0.720	0.787	0.945	00:29
Cognate-Based					
NED	0.853	0.705	0.771	0.941	01:48
SCA	0.862	0.719	0.783	0.944	04:49
Classifier-Based SVM (linear)					
NED, SCA	0.931	0.713	0.806	0.952	00:37

Table 2: Ten-fold cross-validation for three methods with NED (normalized edit) and SCA (Sound-Class based phonetic alignment) distance measures. **Bolded** estimates are superior to and underlined estimates inferior to the the overall average using analysis of means (Nelson et al., 2005) with joint error rate $\alpha = 0.05$.

would perform as well as all languages together. A combined trial performs better; a single threshold estimate appears to generalize better to held-out data than using individual language estimates.

Experiment	Prec.	Rec.	F1	Acc.
Classifier Variations - NED, SCA				
SVM (rbf)	0.945	0.694	0.799	0.951
Logistic regression	0.914	0.728	0.809	0.952
SVM (balanced)	0.613	0.826	0.704	0.902
Closest Match - SCA				
Each language (avg)	0.860	0.707	0.770	0.941

Table 3: Ten-fold cross-validation for several *ad hoc* experiments with NED (normalized edit) and SCA (Sound-Class based phonetic alignment) distance measures. Classifier experiments: SVM with radial basis function, Logistic regression, linear SVM with balanced class weights. Descriptive statistics only.

Tab. 4 shows the results of the Classifier-Based method for the seven target languages in our sample with training and evaluation over the entire dataset. There is some variation in performance by language, in particular, with recall in $[0.615, 0.778]$. We detect a linear relation between the performance and the amount of borrowings from the dominant language in the target languages. (Precision: $r = -0.39, NS$; Recall: $r = 0.88, p < 0.01$; F1 score: $r = 0.85, p < .01$; 1-sided Pearson correlation tests with $df = 5$). Recall and F1 scores improve as borrowing increases. This could be an artifact of higher borrowing resulting in better estimation of a target language coefficient, or more interestingly, a cultural process where more dominant-donor borrowing corresponds to reduced phonetic adaption into the target language.

Language	Prec.	Rec.	F1	Acc.	Borr.
Imb. Quechua	0.921	0.773	0.841	0.924	26%
Mapudungun	0.944	0.716	0.814	0.950	15%
Otomi	0.932	0.692	0.794	0.968	9%
Q'eqchi'	0.934	0.615	0.742	0.961	9%
Wichí	0.952	0.658	0.778	0.953	12%
Yaqui	0.938	0.778	0.851	0.941	22%
Zin. Tzotzil	0.932	0.661	0.773	0.949	13%
Average	0.934	0.714	0.810	0.952	15%

Table 4: Individual language results for the Classifier-Based borrowing detection methods on the seven target languages in our sample. The last column shows the proportion of Spanish borrowings.

ID	DOCULECT	TOKENS	DONOR LANGUAGE	DONOR VALUE	DET STATUS
▼ ABSTAIN FROM FOOD					
	Spanish	a i u n a r			
6227	Qeqchi	a i u : n i n k + r i f	Spanish	avunar	fn
▼ ADOBE					
	Spanish	a ð o ß e			
8988	ImbaburaQuechua	a d u b i	Spanish	adobe	fn
10182	Wichi	a l u l i s	Spanish	adobe	fn
▼ AGE					
	Spanish	e ð a ð			
10531	Wichi	a n i o	Spanish	año	fn
▼ ANIMAL					
	Spanish	a n i m a l			
3351	ZinacantanTzotzil	tʃ a n u l i l			fp

Figure 2: Example collection of detection errors.

To get a better understanding about the different types of errors that our best performing experimental combination commits, we conducted a detailed error analysis from the Classifier-Based borrowing detection results. A spreadsheet snippet (Fig. 2), serves as a reference for several error types. For undetected borrowings (false negatives), we identified four error types: (1) cases where the borrowed form was not present in the donor wordlist, e.g., Mapudungun *peso* “coin” is borrowed from Spanish *peso* “peso”, but our Spanish wordlist only has *moneda*, (2) cases where the form was present in the donor wordlist, but with a different concept, e.g., Wichi *anio* “age” is borrowed from Spanish *año* “year”, while the Spanish word for “age” is *edad*, (3) cases of large phonetic distance between donor and recipient forms, e.g., Wichi *alulis* “adobe”, which is somewhat distant from Spanish *adobe*, and (4) cases of unrecognized partial borrowing, e.g., Qeqchi *aiunink-rif* “abstain from food”, which is partially borrowed from Spanish *ajunar* “fast”. For falsely detected borrowings (false positives), we identified three error types: (1) cases where the form was not borrowed from the dominant language but *vice versa*, e.g., Spanish *poroto* “bean” was borrowed from Quechua *pu-*

Undetected Borrowings		
Error Type	Count	Pct
borrowed form not in donor list	28	17
different concept than recipient form	75	45
large phonetic distance	31	19
partial borrowing as only reason	5	3
Subtotal	139	84
Falsely Detected as Borrowings		
Error Type	Count	Pct
direction not from dominant donor	7	4
chance similarity of form	10	6
likely dataset error	9	5
Subtotal	26	16
Total	165	100

Table 5: Summary over sample of undetected (false negative) and falsely detected (false positive) borrowings.

rutu, (2) cases of chance similarities between word forms, e.g., Spanish *animal* “animal” and Zinacantan Tzotzil *tʃanulil*, and (3) cases so improbably similar that we suspect errors in the original annotation, e.g., Spanish *pelota* “ball” and Wichi *pelutaj*.

For a large sample of concepts, we tallied 139 undetected (false negative) and 26 falsely detected (false positive) borrowings (see Tab. 5). Most errors were in recall, with many of these borrowings from lexemes **not** within the same concept.

5 Conclusion

How well can we automatically detect borrowings from dominant languages based on wordlist data? We devised three general methods to detect borrowed words from dominant languages, two based on sequence comparison workflows and one based on a classifier. The classifier-based method showed the best performance, with F1 scores of 0.81, and high precision of 0.93. This method could already prove very useful in computer-assisted workflows. Our investigation of detection errors shows several opportunities for improvement. Most undetected borrowings result from current methods’ restriction to searching only for word forms for the same concept. We see great potential in improvements that account for borrowing accompanied by *semantic shift*, specifically: (1) augment donor wordlist coverage of possible forms, (2) relax “same” to “similar” concept restriction for matching forms, or (3) fit language models to wordlists and add word cross-entropy to the classifier without restriction. Tests adding word cross-entropy, not reported here, look very promising, but more research is needed.

Limitations

In this study, we apply limited methods for detection of lexical borrowing to the case of a single dominant donor language (Spanish) wordlist versus seven Latin American language wordlists. This application uses relatively sparse data, with an average of 1,512 lexemes per language wordlist, with very few estimated parameters, and so apt for lower resource languages. Future work will seek to remove the wordlist limitation.

Ethics Statement

Our data are taken from publicly available sources. There are no ethical issues or conflicts of interest in this work.

Supplementary Material

The supplementary material accompanying this study contains the data and code needed to replicate the results reported here, along with detailed information on installing and using the software. It is curated on GitHub (<https://github.com/lexibank/sabor>, Version 1.0) and has been archived with Zenodo (<https://doi.org/10.5281/zenodo.7591335>).

Acknowledgements

This research was supported by the Max Planck Society Research Grant *CALC*³ (JML, <https://digling.org/calc/>), the ERC Consolidator Grant *ProDuSemy* (JML, Grant No. 101044282, see <https://cordis.europa.eu/project/id/101044282>), and by the Graduate School of the Pontificia Universidad Católica del Perú (PUCP) through the Huiracocha-2019 scholarship program, see <https://posgrado.pucp.edu.pe> (JEM). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them. We thank the anonymous reviewers for helpful comments and all people who share their data openly, so we can use it in our research.

References

Cormac Anderson, Tiago Tresoldi, Thiago Costa Chacon, Anne-Maria Fehn, Mary Walworth, Robert

Forkel, and Johann-Mattis List. 2018. A Cross-Linguistic Database of Phonetic Transcription Systems. *Yearbook of the Poznań Linguistic Meeting*, 4(1):21–53.

L. Campbell. 2013. *Historical Linguistics: An Introduction*. Edinburgh University Press.

Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker’s guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Robert Forkel and Johann-Mattis List. 2020. Cldfbench. give your cross-linguistic data a lift. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, pages 6997–7004, Luxembourg. European Language Resources Association (ELRA).

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymiski, 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(180205):1–10.

Harald Hammarström, Martin Haspelmath, Robert Forkel, and Sebastiaon Bank. 2022. *Glottolog [Dataset, Version 4.6]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

A Hantgan, H Babiker, and J List. 2022. First steps towards the detection of contact layers in bangime: a multi-disciplinary, computer-assisted approach [version 2; peer review: 2 approved]. *Open Res Europe*, 2:10:1–25.

Martin Haspelmath and Uri Tadmor. 2009. *World Loanword Database*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Gereon A. Kaiping and Marian Klamer. 2022. The dialect chain of the Timor-Alor-Pantar language family. *Language Dynamics and Change*, 2:1–53.

Mary Ritchie Key and Bernard Comrie, editors. 2015. *Intercontinental Dictionary Series (IDS)*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

V. I. Levenshtein. 1965. Dvoičnye kody s ispravleniem vypadenij, vstavok i zameščenij simvolov [binary codes with correction of deletions, insertions and replacements]. *Doklady Akademij Nauk SSSR*, 163(4):845–848.

- Johann-Mattis List. 2012. [Multiple sequence alignment in historical linguistics](#). In *Proceedings of ConSOLE XIX*, pages 241–260.
- Johann-Mattis List. 2019. Automated methods for the investigation of language contact situations, with a focus on lexical borrowing. *Language and Linguistics Compass*, 13(e12355):1–16.
- Johann-Mattis List, Cormac Anderson, Tiago Tresoldi, and Robert Forkel. 2021. *Cross-Linguistic Transcription Systems [Dataset, Version 2.1.0]*. Max Planck Institute for the Science of Human History, Jena.
- Johann-Mattis List and Robert Forkel. 2022a. [Automated identification of borrowings in multilingual wordlists \[version 3; peer review: 4 approved\]](#). *Open Research Europe*, 1(79):1–11.
- Johann-Mattis List and Robert Forkel. 2022b. *LingPy. A Python library for quantitative tasks in historical linguistics [Software Library, Version 2.6.9]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymiski, Johannes Englisch, and Russell D. Gray. 2022a. [Lexibank, a public repository of standardized wordlists with computed phonological and lexical features](#). *Scientific Data*, 9(356):1–16.
- Johann-Mattis List, Simon J. Greenhill, and Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE*, 12(1):1–18.
- Johann-Mattis List, Annika Tjuka and Christoph Rzymiski, Simon J. Greenhill, Nathanael E. Schweikhard, and Robert Forkel. 2022b. *CLLD Concepticon [Dataset, Version 2.6.0]*. Max Planck Institute Evolutionary Anthropology, Leipzig.
- Jürgen M. Meisel. 2018. [Early child second language acquisition: French gender in german children](#). *Bilingualism: Language and Cognition*, 21(4):656–673.
- Phillipe Menecier, John Nerbonne, Evelyne Heyer, and Franz Manni. 2016. [A Central Asian language survey](#). *Language Dynamics and Change*, 6(1):57–98.
- Chenggang Mi, Lei Xie, and Yanning Zhang. 2020. Loanword identification in low-resource languages with minimal supervision. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(3).
- John E. Miller, Tiago Tresoldi, Roberto Zariquiey, César A. Beltrán Castañón, Natalia Morozova, and Johann-Mattis List. 2020. Using lexical language models to detect borrowings in monolingual wordlists. *PLOS One*, 15(12):e0242709.
- Francesca R. Moro, Yunus Sulistyono, and Gereon A. Kaiping. 2023. [Detecting papuan loanwords in alorese: Combining quantitative and qualitative methods](#). In *Traces of Contact in the Lexicon*, pages 213–262. Brill.
- Abhijnan Nath, Sina Mahdipour Saravani, Ibrahim Khebour, Sheikh Mannan, Zihui Li, and Nikhil Krishnaswamy. 2022. A generalized method for automated multilingual loanword detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4996–5013, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- P. R. Nelson, P. S. Wludyka, and K. A. F. Copeland. 2005. The analysis of means: A graphical method for comparing means, rates, and proportions. *SIAM*.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Katharina Prochazka and Gero Vogl. 2017. [Quantifying the driving factors for language shift in a bilingual region](#). *Proceedings of the National Academy of Sciences*, 114(17):4365–4369.
- Christoph Rzymiski, Tiago Tresoldi, Simon Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Arjava, Nataliai Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps, Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell D. Gray, Robert Forkel, and Johann-Mattis List. 2020. [The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies](#). *Scientific Data*, 7(13):1–12.
- SAS Institute Inc. 2021. *Jmp®*, version 16.1.0. Software.
- René van der Ark, Philippe Menecier, John Nerbonne, and Franz Manni. 2007. Preliminary identification of language groups and loan words in Central Asia. In *Proceedings of the RANLP Workshop on Acquisition and Management of Multilingual Lexicons*, pages 13–20.
- Liqin Zhang, Ray Fabri, John Nerbonne, and John Nerbonne. 2021. [Detecting loan words computationally](#). In *Contact Language Library*, pages 269–288. John Benjamins Publishing Company.