LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 14 (2020), pp. 692–702
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24969
Revised Version Received: 8 Nov 2020

KAIPING, GEREON A., OWEN EDWARDS, & MARIAN KLAMER (eds.). 2019. *LexiRumah 3.0.0*. Leiden: Leiden University Centre for Linguistics. (https://lexirumah.model-ling.eu/lexirumah/)

Reviewed by JEROEN WILLEMSEN, *Aarhus University*
YONATAN GOLDSHTEIN, *Aarhus University*

1. Overview[1] LexiRumah (Kaiping, Edwards, & Klamer 2019) is an online database of lexical items from languages spoken in Eastern Indonesia and Timor-Leste, with a particular focus on the Alor archipelago and the islands of Timor, Lamaholot, and Buton. The area houses languages of both Austronesian and Papuan stock, and one of the primary foci of LexiRumah, the Alor-Pantar region, constitutes the westernmost area where speakers of Papuan and Austronesian languages are in contact.[2] At the time of writing, the database contained 117,986 lexical items from 357 language varieties, or *lects*, which includes major and local languages as well as dialects. 279 of these lects belong to the Austronesian language family – the great majority Central-Eastern Malayo-Polynesian and Celebic – and 51 belong to the Timor-Alor-Pantar language family. The other lects are of South Bird's Head, West-Bomberai, East Bird's Head, Hatam-Mansim, Konda-Yahadian, Maybrat, Mor, Mpur, and Tambora stock, each family being represented by between 1 and 12 lects. The data were gathered from 143 sources, roughly half of which are word lists collected specifically for LexiRumah, the other half including dictionaries, grammars, articles, and field notes.

The database is a product of Marian Klamer's NWO Vici project "Reconstructing the past through languages of the present: the Lesser Sunda Islands",[3,4] and provides, among other things, a tool for investigating the history of the Lesser Sunda Islands. Due to a want of both archaeological and genetic research, the Lesser Sunda Islands are severely understudied, and Klamer's project aims to shed light on the region's history through linguistic research. It does so by making available a parallel database with lexical data from a multitude of lects in the region. This database is based on some 600 carefully selected concepts with varying retention rates, i.e. it includes highly stable, basic vocabulary as well as highly borrowable terms (Kaiping & Klamer 2018: 4), enabling investigations into both horizontal and vertical transmission of

[1]We are grateful to Marian Klamer, George Saad, and two anonymous reviewers, all of whom provided suggestions and comments that increased the quality of this review significantly.

[2]The term Austronesian here refers to the branches of Malayo-Polynesian languages spoken in Eastern Indonesia. Papuan refers to any non-Austronesian language spoken in an otherwise Austronesian-speaking area and is made up of a cluster of several language families. It should also be noted that the current version of LexiRumah has a much wider scope than just the Lesser Sundas, which previous versions put a stronger focus on.

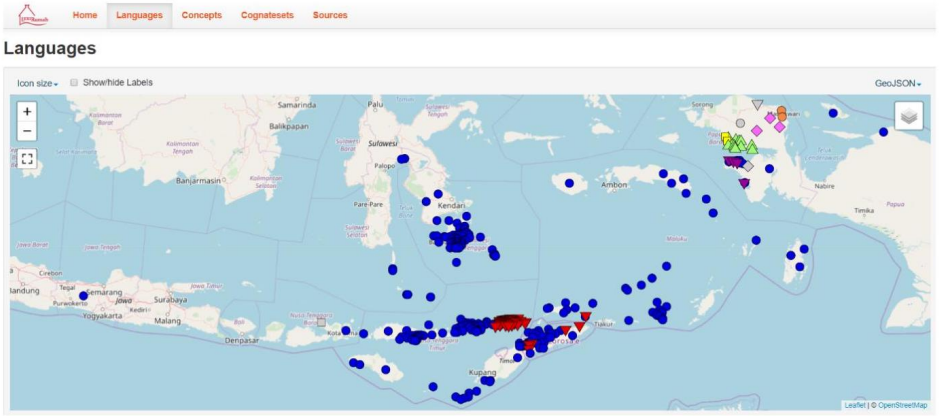
[3]See https://vici.marianklamer.org/.

[4]LexiRumah is part of a three-component database together with GramRumah and CultureRumah, featuring typological and anthropological information respectively, both of which are currently in preparation.

E-ISSN 1934-5275

linguistic features. The patterns of similarity found between lects may then shed light on the type of contact that has taken place between communities. For instance, contact-induced change that is restricted to a small set of lexical borrowings within a particular semantic domain may point to brief contact of a specific nature, whereas linguistic convergence, which is well-attested in Eastern Indonesia as a whole (Klamer, Reesink, & van Staden 2008), may indicate longer, more intensive contact. Further, the on-going inclusion of languages from families outside of the Lesser Sundas enables investigations into wider historical relations (see e.g. Holton & Robinson 2017 on possible links between Timor-Alor-Pantar and other Papuan language families).

Figure 1. LexiRumah's interactive map with different shapes and colours per language family (e.g. Timor-Alor-Pantar is represented by red triangles and Austronesian by blue dots).



LexiRumah is a Cross-Linguistic Linked Database or CLLD (Forkel & Bank 2015). It can be accessed online at https://lexirumah.model-ling.eu/lexirumah/ but can also be downloaded in its entirety along with the Python package PYLEXIRUMAH, which enables the user to manage the data set, e.g. for automated recognition of lexical similarities.[5] Further, individual entries as exposed in the web version can all be downloaded in HTML, JSON, and XSL format. We first discuss the functionality of the online version and the database content and focus on the downloadable data set further below.[6]

2. Web version and database contents The website is structured in a similar fashion to other CLLDs such as WALS (Dryer & Haspelmath 2013), ApiCS (Michaelis et al. 2013), WOLD (Haspelmath & Tadmor 2009) and PHOIBLE (Moran & McCloy 2019), with typical features like interactive maps and filterable information such as

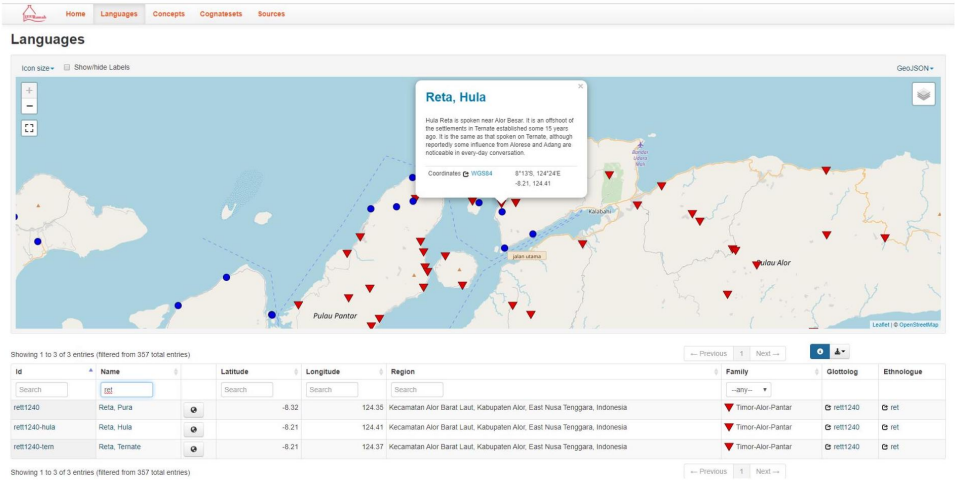
[5]The dataset can be downloaded at https://lexirumah.model-ling.eu/lexirumah/ download, and the dataset plus supplementary material and the Python package is available at https://doi.org/10.5281/zenodo.1164782.

[6]This review does not include a discussion of the Python package. The interactive IPython notebook provided (see Kaiping & Klamer 2018, Supplementary Materials 2) appears to be out of date with Google's current API policy, and no manual is currently available for the package.

language family and geo-coordinates contained within entries. The interface is made up of a homepage with general information about the project, and the database contents are exposed in four tables that provide access to the data from different angles: Languages, Concepts, Cognatesets, and Sources.

In the Languages table, the user is able to browse by lect. The available word forms for a given lect can be accessed either by selecting the lect on the interactive map, which prompts a pop-up with basic information (if available) and a link to the list of available word forms, or directly by clicking on its name in the list below the map. Along with its name, additional data about a given lect is contained in the list, such as geo-coordinates, region, and language family. This information is filterable, allowing the user to narrow down a search according to various parameters. Further, links between lects and external databases – Glottolog and the ISO 639-3 Registration Authority – are given for each lect as well.

Figure 2. Selection of a lect via the list (bottom left) and via the interactive map.



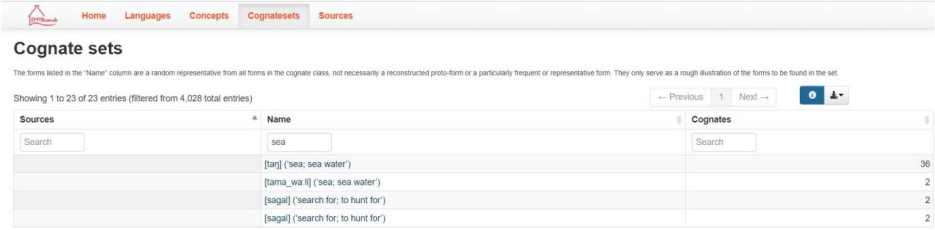
In the Concepts table, the different parameters or concepts are listed, indexed by English simplified glosses. Additional information about these concepts is stored in the list, such as a short Indonesian gloss, the number of lects a concept is available for, its semantic field, and a link to the relevant Concepticon concept set, which can likewise be filtered to narrow down a search. Clicking on a given concept will reveal a list of the lects a concept is available in, again in map and list form. For each lect, the list contains the available word forms both orthographically and in IPA, as well as the source, and, if applicable, a comment on the form (e.g. literal translations, doubts about the accuracy of a form, or analytical comments such as 'inflected for 3SG'). The use of both a practical orthography and IPA makes it suitable for linguistic analyses but also makes it accessible to speakers of the lects themselves.

In the Sources table, the user is able to browse by contribution. The contributions are listed in BibTeX-format and are indexed by a short internal reference. Additional information stored in the list includes the name of the author or fieldworker, the year

the contribution was published or collected, a title, and the BibTeX-type (e.g. 'article', 'phdthesis', 'misc', etc.). The full reference to a given source can be accessed as well, which, in case of fieldnote-based contributions, may contain metadata such as date and location of recording; the consultant's name, age, and (linguistic) background; and the person responsible for recording and transcribing the list.

The Cognatesets table contains observed positive cognate judgments between forms in different lects, as generated by the algorithm LexStat, using the software package LingPy. In the Cognatesets part of the database, each word form (from all included lects) is assigned to a certain similarity class as detected by LexStat, if such a similarity is detected. Each cognate set is represented by an entry in the list that contains the English translation of the concept, as well as a random representative form of the respective reflexes (see Figure 3). This form is neither a reconstructed proto-form nor a form specifically selected because of its representativeness or frequency but rather a random selection from the set of reflexes in the cognate set that helps to identify the set of reflexes for the user. Additionally, the list displays the number of identified reflexes contained in the cognate set. There is also a Source list, presumably meant to link the respective forms in each cognate set to the sources they were gathered from but which is currently still empty. Clicking on a given cognate set reveals the list of forms the set is made up of, per lect (see Figure 4). Also included is an alignment of the sound segments per form to other forms in the same cognate class, providing an indication of which sounds correspond to one another. Sounds that are (presumably) lost or gained in a given form are marked by "-". The label for any given cognate set is composed of a shorthand of the concept and two numbers (e.g. "sea-1-0" in Figure 4).

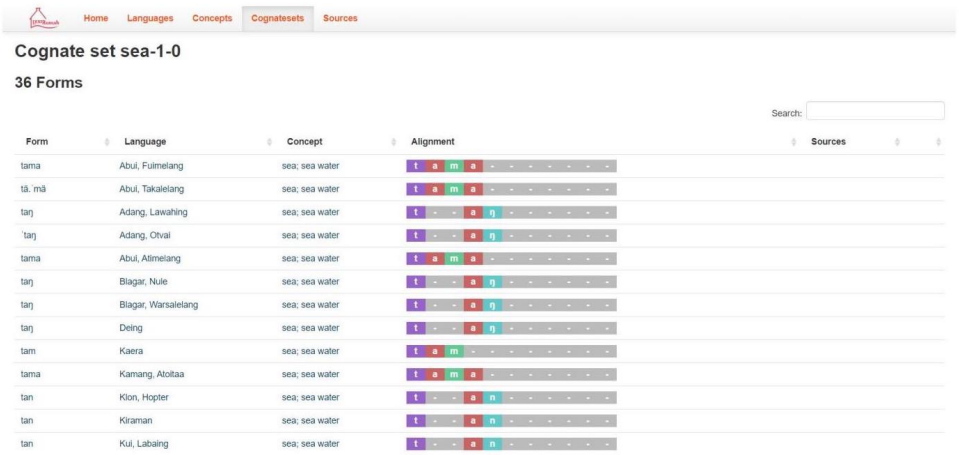
Figure 3. Selection of a cognate set via the list.



The web interface is accessible and easily navigable, and the pages load smoothly and quickly. The different tables and the additional information contained within them are linked in a clear and logical way, making browsing between lects and/or concepts simple and intuitive. The additional information contained within the entry for each lect and word form is made up of enough parameters to narrow down searches without complications, and users interested in a subset of the data, e.g. word forms from a particular language family or area, can easily filter out irrelevant entries.

With 117,986 lexical items from 357 lects, the database has a good coverage of the area. Not only is the multitude of data impressive, but for the most well-represented groups of languages, i.e. the Timor-Alor-Pantar, Central-Eastern Malayo-Polynesian,

Figure 4. A cognate set for the concept 'sea, sea water' including forms, associated lects, and the alignment of sound segments.



and Celebic languages, the coverage also appears to be evenly distributed in relation to the amount of linguistic variety within a languoid as well as the population density of a given area. More lects from densely populated areas (such as Eastern Pantar) are included than from more sparsely populated regions (such as Eastern Alor), and more sources are included for larger languages with several varieties (such as Lamaholot and Keo, with 25 and 16 sources respectively) than for smaller languages with fewer varieties (such as Oirata and Reta, with one and two sources respectively). Another useful feat is the inclusion of seven proto-languages from different branches and levels of the Austronesian and Timor-Alor-Pantar families, providing users interested in comparing concepts in a given lect with the relevant proto-form quick and easy access to both.

A few remarks can be made about the data and the way it is presented, most of which involve minor glitches and inconsistencies, and some of which are mere desiderata. The most significant limitation is perhaps that, given the multitude of data from various types of sources, the quality of the sources obviously varies. While some sources are the outcome of several years of research and/or a longer fieldtrip, some were compiled during a survey, sometimes by a fieldworker otherwise unfamiliar with the language in question. And while certain guidelines were applied to minimise the amount of poor and noisy data, such as consulting small groups rather than individuals, always collecting data in the speaker community, and providing consultants with clear definitions of concepts (Kaiping & Klamer 2018: 14–15), inaccuracy can hardly be avoided in survey-based data collection. This does not pose a major challenge for some types of linguistic analysis, such as large-scale automatic similarity coding, as the multitude of data will ensure that broader patterns of borrowing, retention, and innovation will be revealed even if part of the data is of lesser quality. However, for more fine-grained analyses involving specific word forms, some data may be in-

sufficiently accurate. For instance, the quality of lower-level reconstructions may be compromised if one of the synchronic forms it is based on is inaccurately transcribed.[7]

As for the cognate sets, it would perhaps be advisable to combine the cognate judgements as based on the LexStat analysis with human expertise: some cognates which should be subsumed under the same cognate set are for some reason excluded. The Kula form *tama* 'sea, sea water' is, for instance, not included in the set "sea-1-0" in Figure 4 but forms its own cognate set "sea-1-3". The Cognatesets table would also benefit from a more easily accessible overview of the lects represented in a given set; the representative form (see e.g. *taŋ* in Figure 3) is not always sufficient to recognise them. It is further not entirely clear to the user why the labels for the cognate sets include two numbers. To some users, it might suggest that the sets are divided into subclasses based on relative similarity, which is not the case.

Other issues involve small glitches or inconsistencies and are of minor significance. For example, each lect contains information about the *kecamatan* '~ under-district' in which it is spoken. For some lects, these are outdated, such as Reta (Pura) and Blagar (Pura), which are subsumed under its former *kecamatan*, Kecamatan Alor Barat Laut, rather than the current Kecamatan Pura. Other lects are subsumed under the wrong *kecamatan*, such as Sar (which is spoken in Kecamatan Pantar Timur rather than Kecamatan Pantar), and others still are not subsumed under any. Other minor issues include inconsistencies in the way in which some sources are presented (e.g. 'word list' and 'field notes' used interchangeably) and the inoperative contact form. Lastly, a potential desideratum for some users is the ability to compare entire word lists between sets of languages. Specific word forms can easily be compared across lects, and the entire word list for any given lects is easily retrieved, but in order to compare entire word lists between lects (e.g. to get an idea of the sound correspondences between lects), the user is required to download the data set, which is the topic of the next section.

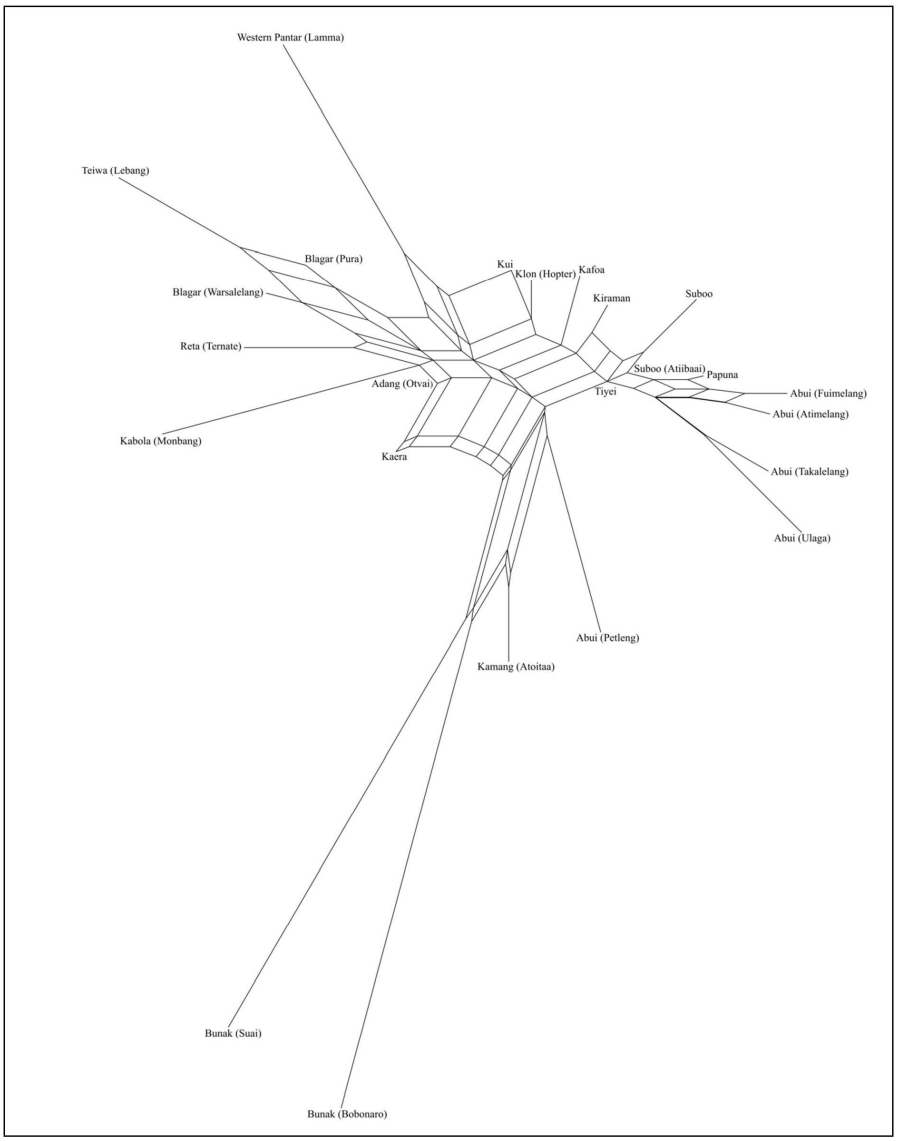
3. The downloadable data set For large-scale data analyses, the entire data set the web interface is based on is downloadable from the website. Here, we demonstrate its usability by showing the results of a distance-based computational analysis, in which we attempted to draw a phylogenetic network using the NeighborNet algorithm as implemented in SplitsTree (Bryant & Moulton 2002; Huson & Bryant 2006) based on the cognate codings of part of the data set.[8]

[7]To provide an example: in Reta and Blagar, which together make up a lower-level branch of the Alor-Pantar languages, the forms for 'shoulder' are *-bɛag* (Reta, Willemsen 2016) and *-abɛa* or *-ɛbɛa* (Pura Blagar, Klamer 2016; Steinhauer & Gomang 2016: 178), respectively, yielding a proto-form along the lines of *-bɛa(g)*. However, one of the current Reta sources mistakenly contains the form *bejag*, based on which this proto-form is not as obvious. Needless to say, the fact that these are cognates would still have been detected by algorithms like LexStat.

[8]The cognate codings for all word forms of a subset of the Timor-Alor-Pantar lects were used to calculate the number of shared cognates, which was then used to draw the network. Using a cut-off point of at least 50 cognates as recognised by LexStat, 25 lects were included. A Python script was used to transform the cognacy codes into a feature matrix listing the respective lects as taxons on the y-axis and cognate codings as binary features on the x-axis. Any cognate was coded with '1', and the absence of any cognate was marked with '0'. If it neither contained the cognate nor another cognate within the same concept, it was treated as a data gap. This resulted in a matrix of 25 languages coded for 1099 features, totaling 27457

Figure 5 below shows the network we managed to produce for a number of Timor-Alor-Pantar languages as an example analysis. The length of any branch represents the lexical distance between two lects, and the 'webbing' in the middle of the network shows the multiple possible routes from one lect to another as a result of conflicting signals in the data.

Figure 5. NeighborNet visualisation of shared cognacy between 25 Timor-Alor-Pantar lects.



data points. A more detailed methodology as well as the Python code can be obtained from the authors at yonatan@cc.au.dk.

Most current conceptions about the classification of Timor-Alor-Pantar languages (Schapper 2017: 8–9) are indeed reflected, which suggests that a multitude of data is indeed able to make up for a certain degree of noise.[9] Further, we found that the data set is easily manipulable due to the accessible Excel-format it is presented in. Another significant plus is the inclusion of a number of Python files and example analyses to further facilitate various kinds of research.

4. Local and non-academic applications While the primary purposes of LexiRumah as a lexical database are largely geared towards scholarly research, care has also been taken to ensure and maximise its usability outside of academia. For one, each lexeme is not only represented by a broad phonetic transcription but also by an orthographical transcription based on Indonesian spelling. For most of the lects included in LexiRumah, there is no written tradition, although some speakers have begun using the lects to communicate through digital media (Saad 2020) largely based on Indonesian orthography, which the great majority of speakers are able to read and write in. This dual representation of lexemes thereby provides the speaker communities with a means to orthographically represent their own language based on a familiar writing system. Thus, speakers are able to use the database as a basic word list to look up words in their local lect as well as to compare words across lects. In addition, this may facilitate grassroots documentation of word forms, stories, linguistic descriptions, etc., but it may also enable both governmental and educational institutions to develop practical orthographies, which may in turn facilitate education in and about local languages.

Furthermore, the inclusion of an interactive map and bibliographic references for each word list also allows for local research into linguistic diversity and language contact. Not only does this type of accessibility facilitate scholarly research within the communities themselves, it may ultimately also work towards bridging the gap between Western and local linguistic research.

Lastly, it can be added that its systematic inclusion and consistent treatment of a great variety of lects also increases its usefulness for speaker communities. Many local languages are under pressure from Indonesian, which is often viewed as a more prestigious language, and local languages may be seen as old-fashioned and inferior (see e.g. Baird 2008: 10–11 on Klon). The consistent treatment of such varieties is invaluable in creating linguistic awareness, both on a local level as well as on a broader governmental or educational level.

5. Conclusion LexiRumah is a large online lexical database of Austronesian and Papuan languages spoken in Eastern Indonesia and Timor-Leste. It makes an in-

[9]For instance, the two Bunak dialects, which are the only lects that are part of the Timor branch, are indeed isolated. The Central Alor lects of Abui, Papuna, Kamang, Suboo, and Tiyei also group together, as do Kui and Kiraman and Klon and Kafoa. The split between the right and left side of the network reflects a split between the Alor lects on the one hand and Straits lects (Reta, Blagar), Pantar lects (Teiwa, Kaera, Western Pantar) on the other. The relative divergence of Western Pantar is also reflected, and the Straits lects indeed form a cluster. The fact that the Straits lects group with Bird's Head lects in some ways and with Pantar lects in others (Schapper 2017: 9) is also reflected in the position of Adang, which is nearly equidistant from the Straits and West-Alor lects.

valuable contribution to the study of the relation between these languages and their speaker communities and the history of the region as a whole. The web-based version of the database is easily navigable and user-friendly and allows the user to browse content from different angles, i.e. lects, concepts, cognate sets, and sources. It is also of direct use for the speaker communities themselves in facilitating local linguistic research and grassroots documentation, as well in increasing linguistic awareness on various levels of society. The main shortcoming of the database content is a certain degree of data noise, which, given the wide scope of the database, is virtually unavoidable. As with much 'big-data'-type research, this noise is cancelled out by the sheer multitude of data, which we briefly demonstrated through an example analysis. Other shortcomings, which mainly consist of inconsistencies in the information included in the database, do not impair its usability in any major way. Overall, we think LexiRumah is an excellent tool for lexical comparison that serves as an example to other cross-linguistic databases.

Primary focus:	Making available large amounts of lexical data from languages spoken in Eastern Indonesia and Timor-Leste for linguistic research and to facilitate comparative research into horizontal and vertical transmission with the goal of shedding light on the region's history.
Pros:	Multitude of (primary) data, logical overall structure, accessible and easy to use, excellent for 'big-data'-type research.
Cons:	Varying accuracy of data, some minor glitches.
Platforms:	Any.
Open source:	Yes, the source code is available at GitHub: https://github.com/lessersunda/lexirumah/.
Proprietary:	Licensed under Creative Commons Attribution 4.0 International.
Available from:	https://lexirumah.model-ling.eu/lexirumah/.
Cost:	None.
Reviewed version:	3.0.0

References

Baird, Louise. 2008. *A grammar of Klon: A non-Austronesian language of Alor, Indonesia*. Canberra: Pacific Linguistics.

Bryant, David & Vincent Moulton. 2002. NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. In Guigó, Roderic & Dan Gusfield (eds.), *Algorithms in bioinformatics*, vol. 2452, 375–391.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (http://wals.info)

Forkel, Robert & Sebastian Bank. 2015. *clld: clld toolkit for cross-linguistic databases*. (https://sandbox.zenodo.org/record/13747)

Haspelmath, Martin & Uri Tadmor. 2009. *World loanword database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (http://wold.clld.org)

Holton, Gary & Laura C. Robinson. 2017. The linguistic position of the Timor-Alor-Pantar languages. In Klamer, Marian (ed.), *The Alor-Pantar languages: History and typology*, 2nd edn., 147–190. Berlin: Language Science Press.

Huson, Daniel H. & David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology & Evolution* 23(2). 254–267.

Kaiping, Gereon A., Owen Edwards, & Marian Klamer (eds.). 2019. *LexiRumah 2.2.3*. Leiden: Leiden University Centre for Linguistics. (https://lexirumah.model-ling.eu/lexirumah/)

Kaiping, Gereon A. & Marian Klamer. 2018. LexiRumah: An online lexical database of the Lesser Sunda Islands. *PLoS ONE* 13(10). 1–29.

Klamer, Marian. 2016. Field notes on Blagar-Pura. In Kaiping, Gereon A., Owen Edwards, & Marian Klamer (eds.). 2019. *LexiRumah 2.2.3*. Leiden: Leiden University Centre for Linguistics. (https://lexirumah.model-ling.eu/lexirumah/)

Klamer, Marian, Ger Reesink, & Miriam van Staden. 2008. East Nusantara as a linguistic area. In Muysken, Pieter (ed.), *From linguistic areas to areal linguistics*, 95–149. Amsterdam: John Benjamins.

Michaelis, Susanne M., Philippe Maurer, Martin Haspelmath, & Magnus Huber (eds.). 2013. *Atlas of pidgin and creole language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (http://apics-online.info)

Moran, Steven & Daniel McCloy (eds.). 2019. *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History. (http://phoible.org)

Saad, George. 2020. *Variation and change in Abui: The impact of Alor Malay on an indigenous language of Indonesia*. Amsterdam: Landelijke Onderzoekschool Taalwetenschap.

Schapper, Antoinette. 2017. Introduction. In Schapper, Antoinette (ed.), *Papuan languages of Timor, Alor and Pantar: Volume 1 – Sketch grammars* (Pacific Linguistics 644), 1–22. Berlin: Mouton de Gruyter.

Steinhauer, Hein. & Hendrik D. R. Gomang. 2016. *Kamus Blagar-Indonesia-Inggris / Blagar-Indonesian-English dictionary*. Jakarta: Yayasan Pustaka Obor Indonesia.

Willemsen, Jeroen. 2016. Field notes on Reta. In Kaiping, Gereon A., Owen Edwards, & Marian Klamer (eds.). 2019. *LexiRumah 2.2.3*. Leiden: Leiden University Centre for Linguistics. (https://lexirumah.model-ling.eu/lexirumah/)

Jeroen Willemsen
jeroen@cc.au.dk

Yonatan Goldshtein
yonatan@cc.au.dk