LA80: A Lexical Database of 10 Bantu A80 Languages

Tessa Y Vermeir, Marc Allassonnière-Tang, Guillaume Segerer

Journal of Open Humanities Data

LA80: A Lexical Database of 10 Bantu A80 Languages

RESEARCH PAPER

TESSA Y. VERMEIR
MARC ALLASSONNIÈRE-TANG
GUILLAUME SEGERER
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

In this paper, we present *LA80*, a database containing lexical data of 10 Bantu A80 languages (Bekwel, Gyeli, Kol, Koonzime, Kwasio, Makaa, Mpiemo, Njyem, Shiwa and Sso). Data from existing fieldwork datasets have been compiled and formatted. We standardised French translations, corrected spelling mistakes, and merged overlapping data points, resulting in a database with 5,588 concepts. Furthermore, for a subset of 557 concepts available in at least six of the 10 languages, we did additional reformatting by separating prefixes from stems, something that is not done systematically in the source data. The LA80 database can be used for comparative linguistic analyses and diachronic reconstructions.

CORRESPONDING AUTHOR:
Tessa Y. Vermeir
DDL, Université Lumière 2/ CNRS, Lyon, France; Éco-Anthropologie, Université Paris-Cité/MNHN/CNRS, Paris, France
t.vermeir@univ-lyon2.fr

KEYWORDS:
lexical database; North-western Bantu languages; corpus analysis; typology; lexical reconstructions

TO CITE THIS ARTICLE:
Vermeir, T., Allassonnière-Tang, M., & Segerer, G. (2024). LA80: A Lexical Database of 10 Bantu A80 Languages. *Journal of Open Humanities Data*, 10: 42, pp. 1–10. DOI: https://doi.org/10.5334/johd.218

(1) INTRODUCTION

The database LA80, Lexicon of A80 Bantu languages, contains lexical data (mostly verbs and nouns but also adverbs, adjectives, and interjections) from 10 related A80[1] languages. Bantu languages of the North-West region are reputed for the variation they present compared to the rest of the family. Where other Bantu languages generally have relatively simple vowel systems, North-western languages often have exceptional inventories (Maddieson & Sands, 2019). Where generally, the Proto-Bantu syllable structure of CV.CV is maintained in the noun stems of current-day languages, North-western languages have created new structures through the loss of segments or syncope, such as word-internal closed syllables (Hyman, 2019). A number of North-western languages no longer have a Low-Low versus Low-High tonal contrast pre-pausally, leading to the Low-High melody being produced as unreleased low tones which may sound like mid tones (Marlo & Odden, 2019). Bantu languages are rich in inflectional and derivational morphology, with the exception of the North-western region (Schadeberg & Bostoen, 2019). A total of 19 noun classes have been reconstructed for Proto-Bantu (Meeussen, 1967, p. 97);[2] out of these 19, a good number are absent in most A80 languages, which generally do not have more than 11 classes (Cheucle, 2014, p. 302). Class 9 has disappeared in some languages because the nouns belonging to this class have been reinterpreted as nouns of class 1, which also has a nasal prefix, and are now of the class pairing 1a/2 (Cheucle, 2014, p. 368).

The LA80 database, which we present in this full-length paper, allows us to verify in a quantitative way qualitatively informed claims or hypotheses. Even though the languages included are very closely related, there is a lot of variation to be investigated, and LA80 provides a typologically interesting small-scale comparative database. Most overview papers on Bantu languages only mention that North-western languages are different or do not fit into the picture (see e.g. Nurse, 2008; Van de Velde et al., 2019, vol. 2, ch. 3–5; Creissels, *To appear* gives a bit more information)—thus much is still to be learned. As more and more data and information become available, it also becomes possible and more informative to do comparative work on these languages.

The paper is structured as follows. Section 2 presents the data included in LA80 and where they came from, as well as the issues that we ran into while constructing the database. In section 3, we show the methodology used to format the database. Section 4 presents a number of ways in which the LA80 database can be used for future research. Some limitations are discussed as well.

(2) COMPILATION OF THE DATABASE

(2.1) DATA SOURCES

The database consists of 10 languages, all from the A80 group. Table 1 shows the languages included, their Guthrie (Cheucle, 2014, p. 14) and Glottolog (Hammarström et al., 2024) codes, and where they are spoken. Cameroon hosts the largest number of languages; only Shiwa is not spoken in this country. Figure 1 shows where the languages are spoken on a map.[3]

The data for eight of the languages come from Marion Cheucle's (2014) thesis; she worked on Bekwel herself and included data on the following seven languages from other sources (see Table 1 for the list of original sources): Shiwa, Kwasio, Makaa, Kol, Njyem, Koonzime and Mpiemo. Cheucle describes in detail how she obtained the data as well as the limitations of her study; the interested reader is referred to the relevant parts of her thesis for more information. The lexical data of all eight languages are available on RefLex (Segerer & Flavier, 2011). The data of the last two languages (Gyeli and Sso) are from two different sources. The data for Gyeli come from Nadine Grimm's (2021) grammar and are also available on RefLex.[4] The data for Sso is first-hand data from Vermeir (forthcoming) and was not yet available on RefLex at the time this paper was written.

1 Languages of the Bantu family are divided into large geographical zones, indicated with a capital letter, and then subdivided into smaller zones, indicated with a two- or three-digit number (Guthrie, 1967; Maho, 2003). This system is purely referential, but it does happen that languages of a subzone are also genetically related, as is the case with the languages we present here (see e.g. Philippson, 2022).

2 In the Bantu tradition, noun classes are numbered. Odd numbers generally indicate a singular form, while even numbers indicate the plural. Odd and even numbers make singular-plural pairs, e.g. a noun can take its singular in class 1 and its plural in class 2, or its singular in class 3 and its plural in class 4, etc.

3 There are no data points for Bekwel and Mpiemo in Cameroon on the Glottolog website.

4 The authors wish to thank Sébastien Flavier for his help extracting the Gyeli data from the RefLex website. It should be noted that these only concern underived forms; the wordlist in Grimm's (2021) also includes derived forms.

LANGUAGE	GUTHRIE CODE	GLOTTOLOG CODE	SPOKEN IN	SOURCE	VARIETY	NUMBER OF WORDS	PLURAL TRANSCR.	NOUN CLASS	TONE TRANSCR.	PHON. TRANSCR.
Gyeli	A801	gyel1242	Cameroon, Equatorial Guinea	Grimm, 2021	Ngolo village (Bulu contact area)	1,495	yes	yes	yes	yes
Shiwa	A803	shiw1234	Gabon	Dougère, 2007	Booué village and area	761	yes	no	no	yes?
Sso	A82	soca1235	Cameroon	Vermeir, forthcoming	"Central"	1,900	yes	yes	yes	yes
Makaa	A83	maka1304	Cameroon	Heath, 1985	Mbwaanz	2,534	yes	yes	yes	yes?
Kol	A832	kolc1235	Cameroon	Henson, 2007	"Central"	1,840	yes	yes	yes	yes?
Kwasio	A84	kwas1243	Cameroon, Equatorial Guinea	Duke, 2004	Mvumbo	1,296	yes	no	no	yes?
Njyem	A84	njye1238	Cameroon, Congo	Beavon & Beavon, 2005	unknown	2,997	no	no	yes	no?
Koonzime	A842	koon1245	Cameroon	Beavon & Beavon, 1996; cited by Cheucle, 2014	Nzime	4,706	yes	yes	yes	yes
Bekwel	A85b	bekw1242	Cameroon, Congo, Gabon	Cheucle, 2014	Ogooué-Ivindo province	2,472	yes	yes	yes	yes
Mpiemo	A86c	mpie1238	Cameroon, Central African Republic, Congo	Beavon & Beavon, 1996–2003; cited by Cheucle, 2014	Bijuki	1,352	no	no	yes	yes?

Table 1 Overview of the 10 languages in the database, their Guthrie and Glottolog codes, the countries they are spoken in, where the data come from and from which variety, the number of words available, and whether the plural forms are included, information on noun classes is included, the tone is transcribed and the transcriptions are phonemic; a question mark indicates that transcriptions may be phonetic.

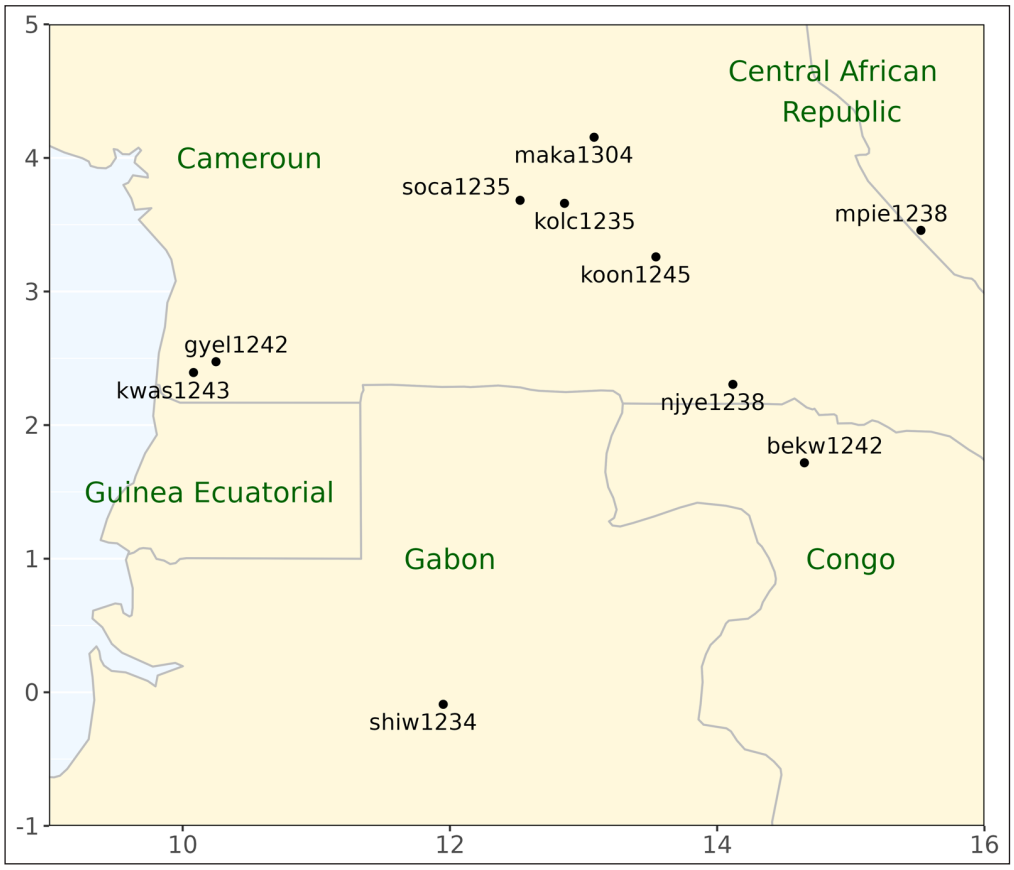

Figure 1 Map indicating where the 10 languages of the database are spoken. The data points are extracted from Glottolog (Hammarström et al., 2024).

The number of words available varies per language, as shown in Table 1. Shiwa has the smallest number of words, not even reaching 1,000. On the other hand, Koonzime has almost 5,000 words and is by far the biggest.

Marion Cheucle (p.c.) has been so kind as to share with us the Excel file in which she had composed her dataset, thus giving us more insight into its constitution and the specifics of each of the lexicons included. However, for the LA80 database, we used only the data made available on RefLex. All the details of each lexicon are presented in Table 1. For the eight languages of Marion Cheucle's dataset, the columns 'Variety' and 'Phonemic transcription' have been completed with information from her thesis (Cheucle, 2014). The information on plural forms, noun classes and tones comes from Marion Cheucle's Excel file (p.c.). All the information on Gyeli comes directly from Grimm (2021), and the information on Sso has been completed by Vermeir (forthcoming), who has done fieldwork on the language.

All of the sources include metadata on the origins of the data, some more precise than others. In half of the cases, the data has been gathered for a thesis; this is the case of Gyeli, Kwasio, Sso, Bekol and Bekwel. The data on Shiwa come from a master's thesis; those on Makaa, Njyem, Koonzime and Mpiemo are descriptions produced outside of an academic context. In half of the sources, the amount of fieldwork is mentioned: 19 months for Grimm (2021), 3 months for Dougère (2007), 9 months in-situ and 6 months online for Vermeir (forthcoming), 18 months for Henson (2007), and 10 months for Cheucle (2014). Sometimes, information on speakers is included as well; Grimm (2021) has worked with up to five speakers in each session, and Dougère (2007), Vermeir (forthcoming) and Cheucle (2014) include speaker profiles in their publications.

As shown in Table 1, some rather important aspects are missing for some languages, i.e. plural forms of nouns, noun classes, and transcription of tone. In the next section, we discuss this issue as well as some others that we encountered while constructing the LA80 database.

(2.2) ENCOUNTERED ISSUES IN THE RAW DATA

Marion Cheucle's (2014) dataset is rather complete and the standardisation of transcriptions has undoubtedly taken a lot of time. However, there are some limitations to the dataset as well.[5] In this section, we will discuss these issues, from what we see as the most important ones to the smaller issues.

The main shortcoming of the dataset is that the words are only included in their full form, meaning that prefixes are not separated from the base forms. This is true for the verbs (which have a prefix marking the infinitive in some of the languages) as well as for the nouns. This makes it difficult to work on stems alone. Since the complete dataset contains almost 18,000 words, going through all of them to manually separate prefixes and stems would take a large amount of time. For the nouns, moreover, it is not always clear what would be the prefix. This is specifically the case when the class prefix consists of only a nasal, since all languages also have prenasalised consonants in their phoneme inventory. For all the words starting with a nasal, in order to know if this nasal is a prefix or part of the stem, it is necessary to know which noun class the word belongs to, information that is also not systematically available. The task is made slightly easier when the plural is available as well, since it regularly happens that the plural prefix replaces the singular one, thus clearly showing which part of the word is the stem. However, this does not work if the prefix of the singular is maintained in the plural form, something that is not uncommon. Moreover, the plural form of nouns is also not always known, as shown in Table 1. These factors all together make it all the much harder to separate prefixes from stems. Marion Cheucle has done the work for the subset of the data for which she has established cognates, but that subset is limited to 1,041 concepts.[6]

Another issue is the information available within the dataset. As already pointed out in the previous paragraph and in section 2, some vital information is missing for a number of languages. For four of the languages, i.e. almost half of the dataset, information on noun classes is missing (Shiwa, Kwasio, Njyem and Mpiemo). For two languages (Njyem and Mpiemo) plural

5 We wish to stress here that these limitations do not stem from Marion Cheucle's work, but are inherent to the data she had access to when she prepared her thesis. She herself lists some limitations in her thesis (Cheucle, 2014, p. 173–174).

6 A concept is e.g. 'dog', and a word is the translation of this concept in a specific language, e.g. 'mpwê' in Njyem or 'mpyó' in Kol.

forms of nouns are also missing; for Kwasio, there is no tonal transcription. This means that for at least three out of 10 languages, only relatively limited data is available. Any typological work on the domain of noun classes is therefore limited.

Formatting is the last issue. Even though Cheucle (2014, p. 174) notes that transcriptions are generally phonemic, she also writes that authors sometimes alternate with phonetic variations. Lastly, it is not uncommon that more than one word is given for a certain concept, without any information as to where the variation comes from (dialectal or inter-speaker for example) nor which word would be the more "standard" form.

The raw dataset as extracted from RefLex and with the addition of the Sso data is made available in the LA80 repository and called LA80-raw.

(2.3) DATABASE METADATA

Repository location

https://doi.org/10.17605/OSF.IO/2P4BN

Repository name

https://osf.io/2p4bn/?view_only=d0fc9a4e794d40f1b388c32c5271ebcc

Object name

LA80

Format names and versions

LA80-main.csv, LA80-long.csv, LA80-subset.csv, LA80-raw.csv

Creation dates

2024-03-18–2024-04-16

Dataset creators

Tessa Vermeir: conceptualisation, curation, resources (laboratoire Dynamique Du Langage, Université Lumière Lyon 2/CNRS, Lyon, France & laboratoire Éco-Anthropologie, Université Paris-Cité/MNHN/CNRS, Paris, France); Marc Allassonnière-Tang: methodology (laboratoire Éco-Anthropologie, Université Paris-Cité/MNHN/CNRS, Paris, France); Guillaume Segerer: resources (laboratoire LLACAN, INALCO/EPHE/CNRS, Villejuif, France)

Language

French, Koonzime, Njyem, Mpiemo, Kol, Shiwa, Kwasio, Makaa, Bekwel, Gyeli, Sso

Licence

Publication date

2024-04-17

(3) FORMATTING OF THE DATABASE: METHODS USED

Besides the raw dataset (LA80-raw, the data extracted from RefLex and the addition of the Sso data), the LA80 database consists of two parts. The LA80-main database is the reformatted data; the LA80-subset is a selection of the LA80-main concepts for which translations in as many languages as possible are available, and on which we did some additional formatting (see below for more details). For comparative work to be informative, around 400 words is a reasonable amount. First, it is considered relatively large when compared with existing lexical databases (Dellert et al., 2020, p. 274–275). Second, less will not capture for example all the phonemes of a language, and more will not drastically increase the quality of the results (see e.g. Dockum & Bowern, 2019). In the remainder of this section, we will describe how we formatted the LA80-main and LA80-subset databases.

As noted, the data for nine of the languages (excluding Sso) are extracted from RefLex in the form of an Excel table. We extracted the following information: ID RefLex, reference, citation (authors), source, unified form, original translation, unified French and English translations (TUF, 'Transcription Unifiée en Français' and TUE, 'Transcription Unifiée en Anglais'), original form, grammatical category, noun class of singular and plural forms, author's comments, compiler's comments, borrowing language (in the case of loan words), derivational forms (for the Gyeli data), and the name of the language. This rendered a table of 18 columns and 19,455 rows. We added the Sso data to this table as well, making a grand total of 21,371 rows.[7] Using an R script, we changed the long format of the table into a wide format by grouping together all the words with the same translation in French: the first column contains the unified French translations (the English translations are excluded at this stage since they are incomplete), followed by the 10 languages of the database. For each language, we also included the grammatical category of the word[8] and noun class information.[9] This resulted in a table of 40 columns and 7,774 rows. Unfortunately, in many cases, a language has more than one word for the same concept in French without any distinction being made. In those cases, we decided to put all the words in the same cell, separated by a semicolon. An example is given in Figure 2.

TUF	konzime_Form	konzime_CGR	konzime_CLS	konzime_CLP
sauce gluante	ámbòl / bàámbòl ; mbòl / mìmbòl	N ; N	1a ; 3	2a ; 4

Figure 2 An example from Koonzime where the language has two words for one concept in French (CGR: grammatical category; CLS: noun class of the singular; CLP: noun class of the plural; N: noun). Singular and plural forms are separated by a slash.

As a next step, we prepared a cleaned-up version of the data, making the following changes. Numerous lines were removed: all the ideophones, expressions and habitual verb forms, all the grammatical morphemes (only included for Kol), all the grammatical words 'translated' in the TUF column with capitals (e.g. 'CONJONCTION'), all the items marked 'X sp.' (for lack of precision), and most of the agentive and action nouns derived from verbs.[10] In many other cases, the content of multiple lines was merged, reducing the total number of rows, and French concepts with the same meaning were grouped together. For example, it often happened that the original translation in the TUF column contained two concepts, 'X; Y', while 'X' and 'Y' also existed separately. In most of these cases, the 'X; Y' row could be split up, i.e. its words could be added to the rows 'X' and 'Y', so that 'X; Y' could be removed. An example is given in Figure 3.

TUF	Change	Gyeli
amour		NA
désir		NA
amour ; désir		kwàlɛ̀ / be-kwàlɛ̀

>

TUF	Change	Gyeli
amour	TRUE	kwàlɛ̀ / be-kwàlɛ̀
désir	TRUE	kwàlɛ̀ / be-kwàlɛ̀

Figure 3 Example of one type of change made to the raw dataset to create a comprehensible database.

Other merges were made when for example the French contained a spelling mistake (e.g. 'cloture' instead of 'clôture'); one row contained a singular form and the other a plural form of the same French concept (e.g. 'conseil' and 'conseils'), and when two concepts were quasi identical (e.g. 'bout de la maison' and 'bout d'une maison'). In a number of cases, the French translation was changed (indicated with 'TUF' in the 'change' column). This was done when the French translation was an inflected verb form, e.g. 'blessé' ('hurt'), while in the raw data, the original translation was 'être blessé' ('to be hurt'); in these cases, we used the original translation. In some other cases, a language had multiple words for the same concept in French, and it turned out that a more precise translation was included in the raw data. However, these changes have not been made consistently, because of the size of the database. Every time a change has been made, this was marked 'TRUE' in an additional column.

7 Since the Sso data is not yet on RefLex at the time of writing this paper, not all columns are filled here; missing columns are ID RefLex, reference, original translation and complete form.

8 Possible values are: adjective (adj), adverb (adv), conjunction (conj), exclamation (exp), interrogative (int), interjective (interj), noun (N), not applicable (NA), numeral (num), preposition (prep), quantitative (qnt), locative (loc), temporal (temp), verb (V), auxiliary verb (Vaux).

9 The other categories that were originally extracted from RefLex have been discarded at this point in order to keep only the most essential information, in an effort to make the LA80-main database as user-friendly as possible. All this information is still available in the raw version of the database (LA80-raw).

10 All these words can still be found in LA80-raw, also included in the repository.

After the reformatting, LA80-main consists of 5,885 rows (instead of 7,774). The number of cells containing more than one word in the 10 languages has diminished from 4,936 to 3,276; the number of cells with more than one concept in French has diminished from 986 to 483.

As noted in the previous section, one of the main issues with the raw data is that the prefixes are not separated from the stems. We selected a subset of the LA80-main database for which we manually separated prefixes and stems. Since we are also interested in cluster analysis (Vermeir & Allassonnière-Tang, forthcoming), the concepts with a translation in as many languages as possible are of main interest here.[11] However, we excluded the cells that contained more than one word, to avoid having to make a choice between them. This yielded a total of 556 concepts available in at least six languages.[12] Table 2 gives an overview of how many concepts are available in how many languages; Table 3 shows how many concepts there are per language.

Number of concepts	32	77	104	159	184
Number of languages	10	9	8	7	6

Table 2 Overview of how many concepts are available in how many languages in the LA80-subset database.

LG	BE-KWEL	GYELI	KOON-ZIME	KOL	KWA-SIO	MAKAA	MPIE-MO	NJYEM	SHIWA	SSO
Nº	458	449	402	454	449	351	358	392	263	485

Table 3 Number of concepts available per language in the LA80-subset database.

The final step concerned the cleaning up of the selected data. We manually separated prefixes from stems by means of an n-dash ('-');[13] the dashes already used in compounds were replaced by a space. For more readability of the data, we put the symbol for nasality underneath the grapheme instead of on top (e.g. 'o̱' instead of 'õ'), where it interfered with tone transcriptions. Reduplication is indicated with a tilde ('~', e.g. Bekwel 'kɥé~kɥén' 'star'). Lastly, we added data on reconstructions from the Bantu Lexical Reconstruction 3 (Bastin et al., 2002) database.[14] Figure 4 shows an example of what the LA80-subset database looks like.

TUF	change	koonzime_Form	koonzime_CGR	koonzime_CLS	koonzime_CLP
abattre	TRUE	è-cwɛ̀l	V	NA	NA
abeille	TRUE	á̱ŋʊ̄kwàn / bà-á̱ŋʊ̄kwàn	N	1a	2a

Figure 4 An example of Koonzime showing what the LA80-subset looks like after the clean-up.

(4) APPLICATIONS OF THE DATABASE AND REMAINING LIMITATIONS

Compared to the RefLex website, the main advantage of the LA80-database is that the data of all 10 A80 languages can be easily compared: one can see right away, for any given concept, which languages have a translation for this concept.

One way in which the LA80 database can be used is for synchronous comparison. Especially the LA80-subset that we have cleaned up is of great interest for comparative work, since stems can be compared separately from prefixes. Even though information on noun classes is missing for some of the languages (see previous section), it will still be informative to compare the forms of the nominal prefixes across the 10 languages. Moreover, the size of the LA80 subset is sufficient for phonological comparison (see e.g. Dockum & Bowern, 2019). This has been done for eight of the 10 languages (Cheucle, 2014), but now Sso and Gyeli can also be included in the comparison. Comparative work on tone is facilitated as well. Even though in the main LA80 database the prefixes are not separated from stems, we have made a serious effort to standardise translations, which makes concepts more comparable across these 10 languages.

11 Cluster analysis groups items together based on how similar they are. In our case, these items would be the languages of the LA80 database, and their similarity would be measured using their vocabulary. The more data points (i.e., words in this case), the better the analysis.

12 There were 419 concepts available in at least six languages before the clean-up of the raw data, out of which only 10 available in all 10 languages.

13 For Gyeli and Sso, this had already been done in the source material by the respective authors (Grimm, 2021; Vermeir, forthcoming).

14 We added reconstructions made for the languages of zone A, without taking into account whether the current-day forms could be reflexes of these reconstructions.

Another way in which LA80 can be used is for reconstruction. Languages of the North-western Bantu area are known to have suffered segmental loss, especially word-finally, which has consequences on the structure of words (Cheucle, 2014). For example, many languages nowadays have closed syllables, while the basic structure for Bantu languages is open syllables. Another consequence is the creation of long vowels or the appearance of contour tones. Comparing a group of closely related languages allows us to reconstruct what the protolanguage looked like and the ways in which languages develop over time (see e.g. Philippson, 2022). Marion Cheucle (2014) has proposed 1,041 reconstructions based on her data;[15] she made a reconstruction for each concept that is a cognate in at least three languages. We are currently working on updating these cognates with the data from Gyeli and Sso (Vermeir, forthcoming). With the addition of these two languages to the corpus, there are now potentially many more concepts that have cognates in at least three languages, and therefore more reconstructions to propose for the proto-A80 language.

Finally, the LA80 subset can also be used by researchers interested in e.g. the prefixes of only one of the 10 languages. Since prefixes and stems are systematically separated, one can easily find the different prefixes of a given language.

The main limitation of LA80-main remains that prefixes are not all separated from stems. In order to use the full potential of the database, this would need to be done systematically. For now, we have only been able to do so for part of the data. Another limitation is that the original datasets vary in quality: sometimes the plural form of nouns is missing, sometimes tone, and sometimes phonemic and phonetic transcriptions are used interchangeably. Ideally, these issues would be resolved through further research. A last limitation is that in quite a number of cases, there is more than one word per language to translate a given concept in French, and it has not always been possible to separate these. Finally, the addition of unified English translations and reconstructed forms from the Bantu Lexical Reconstruction database (Bastin et al., 2002) would make the LA80 database even more complete.

However, we do believe that already in this format, the LA80 database will yield interesting and concrete results when used for further research, and will enlarge our understanding of the ways in which languages structure their lexicons.

ACKNOWLEDGEMENTS

The authors wish to thank Marion Cheucle for sharing the raw data of her thesis with us; Sébastien Flavier for his help extracting the data from RefLex; Rémi Anselme for his help with the R script; and François Pellegrino for his helpful comments on the structure of the LA80 database. All mistakes are our own.

FUNDING INFORMATION

The authors are thankful for the support from the French National Research Agency (ANR-20-CE27-0021).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR INFORMATIONS

Vermeir: conceptualization; data curation; formal analysis; methodology; resources; writing–original draft

Allassonnière-Tang: methodology; software; supervision; visualisation; writing–review & editing

Segerer: resources; writing–review & editing

15 Available for consultation on RefLex at https://reflex.cnrs.fr/Africa/index.php?state=lexical (accessed 27/03/2024).

AUTHOR AFFILIATIONS

Tessa Y. Vermeir orcid.org/0009-0009-1635-099X
DDL, Université Lumière 2/CNRS, Lyon, France; Éco-Anthropologie, Université Paris-Cité/MNHN/CNRS, Paris, France

Marc Allassonnière-Tang orcid.org/0000-0002-9057-642X
DDL, Université Lumière 2/CNRS, Lyon, France; Éco-Anthropologie, Université Paris-Cité/MNHN/CNRS, Paris, France

Guillaume Segerer orcid.org/0000-0003-3712-865X
LLACAN, INALCO/EPHE/CNRS, Villejuif, France

REFERENCES

Bastin, Y., Coupez, A., Mumba, E., & **Schadeberg, T. C.** (Eds.). (2002). *Bantu lexical reconstructions 3 / Reconstructions lexicales bantoues 3* https://www.africamuseum.be/en/research/discover/human_sciences/culture_society/blr

Beavon, K., & **Beavon, M.** (2005). *Précis d'orthographe pour la langue Nyjem*. SIL Cameroon.

Cheucle, M. (2014). *Étude comparative des langues makaa-njem (bantu A80) : Phonologie, morphologie, lexique* (Doctoral dissertation, Université Lumière Lyon 2, France).

Creissels, D. (To appear). Bantu languages: Typology and variation. In E. Hurst, N. C. Kula, L. Marten & J. Zeller (Eds.), *The Oxford Guide to the Bantu Languages*. Oxford University Press.

Dellert, J., Daneyko, T., Münch, A., Ladygina, A., Buch, A., Clarius, N., Grigorjew, I., Balabel, M., Boga, H. I., Baysarova, Z., Mühlenbernd, R., Wahle, J., & **Jäger, G.** (2020). NorthEuraLex : A wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation*, *54*(1), 273–301. DOI: https://doi.org/10.1007/s10579-019-09480-6

Dockum, R., & **Bowern, C.** (2019). Swadesh lists are not long enough: Drawing phonological generalizations from limited data. *Language Documentation and Description*, *16*. DOI: https://doi.org/10.25894/ldd112

Dougère, L. (2007). *Première approche phonologique, morpho-syntaxique et diachronique du chiwa du Gabon (Ogooué-Ivindo)* (Master's thesis,Université Lumière Lyon 2, France).

Duke, D. (2004). *Lexique kwasio*. SIL Cameroon.

Grimm, N. (2021). A grammar of Gyeli (Comprehensive Grammar Library 2). Language science press.

Guthrie, M. (1967). *Comparative Bantu. An introduction to the comparative linguistics and prehistory of the Bantu languages*. Gregg Press LTD.

Hammarström, H., Forkel, R., Haspelmath, M., & **Bank, S.** (2024). *Glottolog 5.0*. Max Planck Institute for Evolutionary Anthropology; Link: http://glottolog.org. DOI: https://doi.org/10.5281/zenodo.10804357

Heath, D. (1985). *Lexique provisoire de 2800 mots de la langue məkaa*. SIL Cameroon.

Henson, B. (2007). *Kol phonology and morphosyntax* (Dissertation,University of California, Berkeley, California).

Hyman, L. M. (2019). Segmental phonology. In M. Van de Velde, K. Bostoen, D. Nurse, & G. Philippson (Eds.), *The Bantu Languages* (p. 128–149). Routledge. DOI: https://doi.org/10.4324/9781315755946-4

Maddieson, I., & **Sands, B.** (2019). The sounds of the Bantu languages. In M. Van de Velde, K. Bostoen, D. Nurse & G. Philippson (Eds.), *The Bantu Languages* (p. 79–127). Routledge. DOI: https://doi.org/10.4324/9781315755946-3

Maho, J. (2003). A classification of the Bantu languages: An update of Guthrie's referential system. In M. Van de Velde, K. A. G. Bostoen, D. Nurse & G. Philippson (Eds.), *The Bantu languages* (p. 639–651). Routledge. DOI: https://doi.org/10.4324/9781315755946-5

Marlo, M. R., & **Odden, D.** (2019). Tone. In M. Van de Velde, K. A. G. Bostoen, D. Nurse & G. Philippson (Eds.), *The Bantu languages* (p. 150–171). Routledge.

Meeussen, A. E. (1967). Bantu grammatical reconstructions. *Africana Linguistica*, *3*, 79–121. DOI: https://doi.org/10.3406/aflin.1967.873

Nurse, D. (2008). *Tense and aspect in Bantu*. Oxford University Press. DOI: https://doi.org/10.1093/oso/9780199239290.001.0001

Philippson, G. (2022). Double reflexes in north-western Bantu and their implications for the Proto-Bantu consonant system. In K. Bostoen, G.-M. de Schryver, R. Guérois & S. Pacchiarotti (Eds.), *On reconstructing Proto-Bantu grammar* (p. 3–58). Language science press. https://library.oapen.org/bitstream/handle/20.500.12657/63273/external_content.pdf?sequence=1#page=55

Schadeberg, T. C., & **Bostoen, K.** (2019). Word formation. In M. Van de Velde, K. Bostoen, D. Nurse & G. Philippson (Eds.), *The Bantu Languages* (p. 172–203). Routledge. DOI: https://doi.org/10.4324/9781315755946-6

Segerer, G., & **Flavier, S.** (2011). *RefLex: Reference Lexicon, Version 2.2*. https://reflex.cnrs.fr

Van de Velde, M., Bostoen, K. A. G., Nurse, D., & **Philippson, G.** (Eds.) (2019). *The Bantu languages* (Second edition). Routledge. DOI: https://doi.org/10.4324/9781315755946

Vermeir, T. (forthcoming). Lexical and phonological analysis of Sso and related A80 Bantu languages. Université Lumière Lyon 2.

Vermeir, T., & **Allassonnière-Tang, M.** (forthcoming). Cluster analysis on ten related Bantu A80 languages.

TO CITE THIS ARTICLE:
Vermeir, T., Allassonnière-Tang, M., & Segerer, G. (2024). LA80: A Lexical Database of 10 Bantu A80 Languages. *Journal of Open Humanities Data*, 10: 42, pp. 1–10. DOI: https://doi.org/10.5334/johd.218

Submitted: 17 April 2024
Accepted: 29 May 2024
Published: 08 July 2024

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.