Article

Nicolas Quint and Marc Allassonnière-Tang*

Inferring case paradigms in Koalib with computational classifiers

https://doi.org/10.1515/cllt-2021-0028
Received April 24, 2021; accepted December 16, 2021; published online January 20, 2022

Abstract: The object case inflection in Koalib (Niger-Congo) represents complex patterns that involve phoneme position, syllable structure, and tonal pattern. Few attempts have been made with qualitative and quantitative approaches to identify the rules of the object case paradigms in Koalib. In the current study, information on phonemes, tones, and syllables are automatically extracted from a Koalib sample of 2,677 lexemes. The data is then fed to decision-tree-based classifiers to predict the object case paradigms and extract the interactive patterns between the variables. The results improve the predicting accuracy of existing studies and identify the case paradigms predicted by linguistic hypotheses. New case paradigms are also found by the computational classifiers and explained from a linguistic perspective. Our work demonstrates that the combination of linguistic theoretical knowledge with machine learning techniques can become one of the methodological approaches for linguistic analyses.

Keywords: decision trees; Koalib; object case; rules; tone

1 Introduction

The main goal of this paper is to identify (i) the main categories of object case paradigms in Koalib (ii) the information relevant for predicting those categories. The Koalib language has 100,000 speakers that are mainly found in the Nuba Mountains, which cover most of the Province of South Kordofan, in the Republic of Sudan (Quint 2010a). The language is generally affiliated to the Heibanic (or Heibanian) group of the Kordofanian branch of the Niger-Congo language family

***Corresponding author: Marc Allassonnière-Tang**, EA UMR7206, MNHN/CNRS/Université de Paris, Paris, France, E-mail: marc.allassonniere-tang@mnhn.fr. https://orcid.org/0000-0002-9057-642X
Nicolas Quint, LLACAN UMR8135, CNRS/INALCO/EPHE, Paris/Villejuif, France, E-mail: nicolas.quint@cnrs.fr

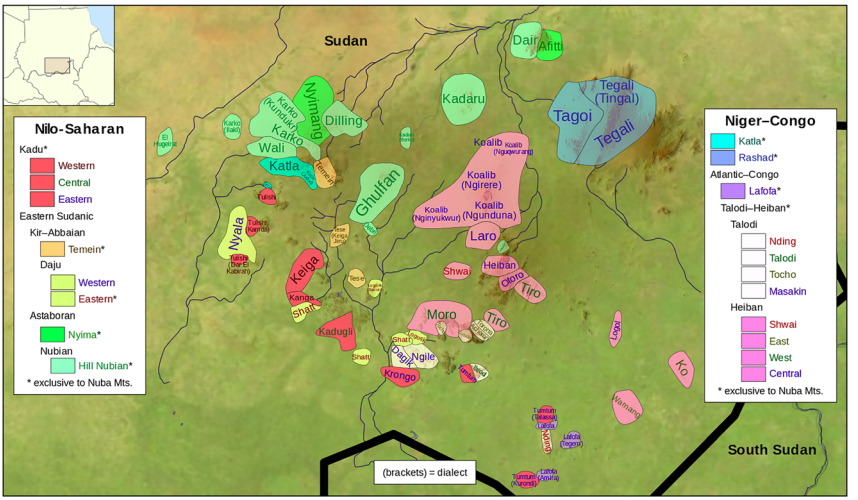

Figure 1: A geographical overview of Koalib and its neighbor languages in the Nuba Mountains (Sudan).

(Quint 2020; Schadeberg 1981), which is the biggest language family in the world in terms of individual languages.[1] This study focuses on the Rere variety of Koalib, spoken in the center of the Koalib linguistic area and considered as the basis of written modern Koalib. The linguistic data available for Koalib and the surrounding languages are still quite limited due to different factors, the most significant of which are a difficult access to the field, a high number of typologically idiosyncratic features, and a considerable linguistic diversity (Dimmendaal 2015; Quint 2006, 2020). Indeed the Nuba Mountains qualify as one of the most important hotspots for linguistic diversity: as shown in Figure 1,[2] in the Nuba Mountains, alongside Koalib and other Heibanian languages, many more languages are spoken in this small geographical area and it is common for each language to have several dialects, rendering more challenging the task for descriptivists. Moreover, the structure of a Kordofanian language such as Koalib is rather complex due to the extensive use of tones (especially in morphology), vowel harmony, noun classes, and a rich inflectional morphology (Quint and Ali Karmal Kokko 2009). Thus, a computer-assisted approach could provide additional insight to reveal hidden patterns of the language system.

1 The affiliation of Kordofanian to Niger-Congo is not uncontroversial in the literature, as the lack of large-scale and consistent linguistic data may have induced chance resemblance (Hammarström 2013; Hammarström et al. 2019).

2 https://commons.wikimedia.org/wiki/File:Map_of_the_languages_of_the_Nuba_Mountains.svg.

The Koalib tonal system distinguishes two register tones and two main contour tones. The register tones are low (L) and high (H), whereas the most common contour tones are falling (F) and raising (R) (Quint and Ali Karmal Kokko 2009). A F tone starts as H and ends in L and a R tone starts as L and ends in H. L and H are the default tones, whereas contour tones are quite rare in Koalib due to their more complex pattern. In the current study, tones are indicated by accents marked above vowels, c.f., [à] for L, [á] for H, [â] for F, and [ă] for R. Vowel harmony implies that vowels in a given word can only belong to one and the same set. Two vowel harmonic sets can be identified based on height (and possibly on tongue root position): (i) a high set comprising with /i/, /ɐ/, and /u/ and (ii) a low set comprising with /e/, /ɛ/, /a/, /ɔ/, and /o/. Finally, Koalib also uses a grammatical system of noun classes, which functions in a similar way as the grammatical gender system found in Indo-European languages (Corbett 1991, 2013). Nouns in Koalib are affiliated to 13 different noun classes according to different semantic criteria (Quint 2013, 2022). For instance, class /ŋ/ mostly refers to liquids, class /kw/ (PL /l/) to human beings, class /t/ (PL /r/) to long objects etc. The affiliation of a noun to a specific noun class triggers grammatical agreement in other elements of the clause.

Koalib also has a nominal declension, with two cases, subject and object. The subject case is the unmarked case in Koalib, whereas the object case (i.e., the accusative case) is either also unmarked or indicated by a suffix and/or a change of the tonal pattern. As an example, in (1a), 'sheep' is the subject of the clause; it is unmarked and codes for the subject case. In (1b), 'sheep' is the object of the clause and undergoes a change in form and tonal pattern: the suffix [è] is added and the tonal pattern changes from HL to LHL, with an additional toned syllable on the suffix.

(1) Example of object-case inflection in Koalib

a. *káaŋàl* *ŋkó* *kè-pèetò*
sheep.S DEM.CLF.PROX CLF-be.white.PFV
'This sheep is white'

b. *Kwókkò* *kwèm-ɛ̀ɛcɛ́* *kàaŋálè* *ŋkó*
Kwókkò CLF.PRF-see sheep.O DEM.CLF.PROX
'Kwókkò has seen this sheep'

The object case inflectional paradigm is generally considered hard to identify since it involves an interaction of several features mentioned below. Four main types of object case inflection can be found: (i) 'same form between the subject and object case' (ii) 'suffixation' (iii) 'tone change' (i.e. change of tonal pattern) (iv) 'both suffixation and change of tonal pattern'. Table 1 shows an overview of the four

Table 1: Main types of object-case inflection in Koalib. The data and examples are from a sample of 1,200 nouns in Koalib (Boychev 2013, p. 8).

Type	Ratio	Subject form	Object form
Same form	25%	ŋèráaɽà (sauce)	ŋèráaɽà (sauce)
Suffixation	35%	kòʈʈó (gourd)	kòʈʈóŋé (gourd)
Tone	6%	kwìcì (person)	kwícì (person)
Both	34%	kwòtlòm (jackal)	kwótlòmá (jackal)

types in terms of ratio based on observations in previous studies (Boychev 2013; Quint 2010b). As an example, the change from *kwìcì* to *kwícì* does not involve any segmental modification, but its tonal pattern changes from LL (subject) to HL (object). The noun is thus labeled with the case paradigm 'tone change'. In terms of distribution, a small fraction of the nouns only undergo a tone change, whereas the three other types display a similar ratio.

Previous studies investigated the object case paradigms resorting to both qualitative and quantitative approaches. Qualitatively, semantically motivated markers are scarce but easily identifiable. For instance, the object case of proper nouns is consistently realized with the /ŋwó/ suffix. Moreover, the high number of parameters that must be taken into account in order to correctly predict the object form of a given noun renders it quite difficult to find the rules for a human brain. That is the reason why the first author of this paper (a descriptive linguist) began quite early to collaborate with computer linguists, such as Boychev. Quantitatively, rule-based classifiers have been used as an attempt to generate the object case paradigms in Koalib (Boychev 2013). These rule-based classifiers consider suffixation and tone change separately. Their average reported accuracy is 66% on a dataset of 1,200 Koalib nouns with a majority baseline of 30%. We further develop these analyses by extending the use of computational methods and expanding the linguistic analysis of the results.

2 Materials and methods

The main source of data in this study is a Koalib corpus collected by the first author, who has conducted fieldwork on the language for more than 20 years, including 15 months of immersion among Koalib speakers and 13 more months spent with Koalib speakers outside the Koalib community. The corpus includes conversations and narratives (approximately 2 h 30 min of recordings, which have been transcribed and aligned), as well as elicitations and the systematical scrutiny of several

books (The New Testament 1967, 1993), pedagogical materials (Abdalla and Komi 2000; Abdalla Omer et al. 1995, 1998), and stories (Karshola Omar et al. 2000; Kodi 2000; Suliman 2000) published in Koalib. These publications represent in all a corpus of 600,000 words. We do not list the detailed ratio of genres and/or registers since the current study focuses on case marking, which is not subject to change under different genres and/or registers in Koalib. Furthermore, we acknowledge that this Koalib corpus is relatively small compared to languages with a large amount of data available (such as English). Nevertheless, we point out once again that this is precisely one of the challenges for working on less-documented languages. Finally, to further enhance the data available for the current analysis, while the paradigms that we are investigating here were mostly extracted from this corpus, a minority of missing cells were filled through elicitation sessions. We consider that this process of elicitation is not likely to introduce biases in the data, since the current study analyses case marking paradigms without considering the discourse frequency of each individual noun. Note also that this corpus is not yet publicly available as its content is in course of publication (Quint and Ali Karmal Kokko 2022).

Information on 2,677 nouns is extracted from the corpus. The data includes information on the lexeme, case, etymology, noun type (e.g. deverbal, proper noun), noun class, and animacy. The lexeme refers to the unmarked subject form (Quint and Ali Karmal Kokko 2009). The case refers to the noun form in the object case, e.g. the entry *cónʈàŋ* 'lion' (subject) is associated with the object form *còn-ʈáŋè*. Both forms are encoded in a mostly IPA-based phonological orthography (pp. 189–210, 34; pp. 169–187, 41). The data is thus considered as a realistic phonological representation of the nouns. The etymology indicates if the noun is a loanword from foreign languages such as English (e.g. *képèn* 'shroud' < English *coffin*) or Arabic (e.g. *àrcâc* 'bullet' < Arabic *raṣáaṣ* 'lead (metal)' (for more details about lexical borrowings in Koalib, see 37; 38). The noun type differentiates proper nouns (e.g. *Kwókkò* 'name of the first-born male child'), common nouns (e.g. *ŋâo* 'water'), and deverbals (e.g. *táakà* 'marriage' < *àaké* 'marry (a woman)'). The noun class indicates the affiliation of the noun to a specific noun class. Finally, information about the animacy of the referent of the noun is also provided, e.g., *kwór* 'man' is counted as animate while *kâl* 'stone' is counted as inanimate. Figure 2 shows the distribution of each value in the different categories annotated in the data.

Most of the nouns are common nouns (69%, 1,830/2,677), with 20% (546/2,677) being proper nouns and 11% (300/2,677) being deverbals. In terms of etymology, the majority of the nouns are affiliated to a Koalib source (84%, 2,253/2,677), while most of the borrowed nouns can be traced back to the Arabic language (15%, 393/2,677). Only a small portion of nouns (1%, 31/2,677) comes from

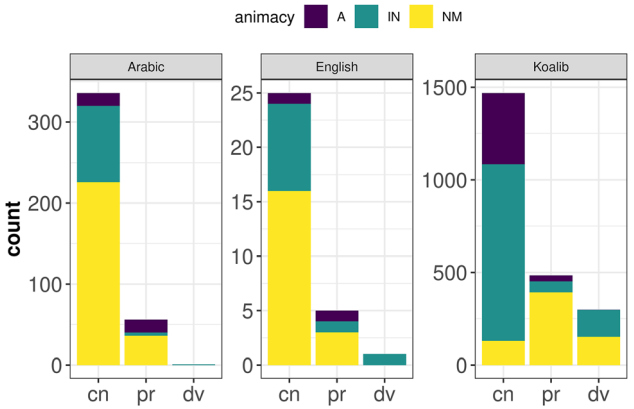

Figure 2: The distribution of main annotated categories in the data. The facets refer to the etymology of the nouns. The abbreviations are interpreted as follows: A = animate, IN = inanimate, NM = not mentioned, cn = common nouns, pr = proper nouns, dv = deverbals.

English. In terms of animacy, half of the nouns are identified as inanimates (47%, 1,268/2,677), 17% (453/2,677) of the nouns are labeled as referring to animates, while this status has not been specified for the remaining 36% (956/2,677).

2.1 Extracting information from the data

Information on phonemes and tones is extracted automatically for the change from subject forms to object forms. Table 2 shows a sample of the encoded data. The first two columns display the raw forms of the subject and object forms.[3] The third column encoding the meaning of the words is shown in the current table in order to ease reading of the paper. However, it is not included in the actual data. The fourth column indicates the segmental changes, e.g., adding a suffix /a/ or /e/. The fifth column indicates if a final toned syllable is added to the original tonal structure, e.g., in the first row, the subject tonal scheme HB changes to HBH in the object case. A final tone H is added. The sixth column marks the number of syllables on the subject form of the word. The seventh column indicates if a change of tonal pattern

3 For practical purposes, the subject and object forms have been disharmonized in the database, i. e. the high vowels /i, ɐ, u/ (see above Section 1) are represented like their low counterparts /e, a, o/. This is motivated by the fact that there is no contrast between high and low vowels in the same Koalib word and harmony does not seem to play any role in case inflection. For instance, the Koalib equivalent of English 'termitary', represented phonologically as /ɛ̀ɐ̀tùm/ in the data, is coded in our database as *àatòm* (see Table 4).

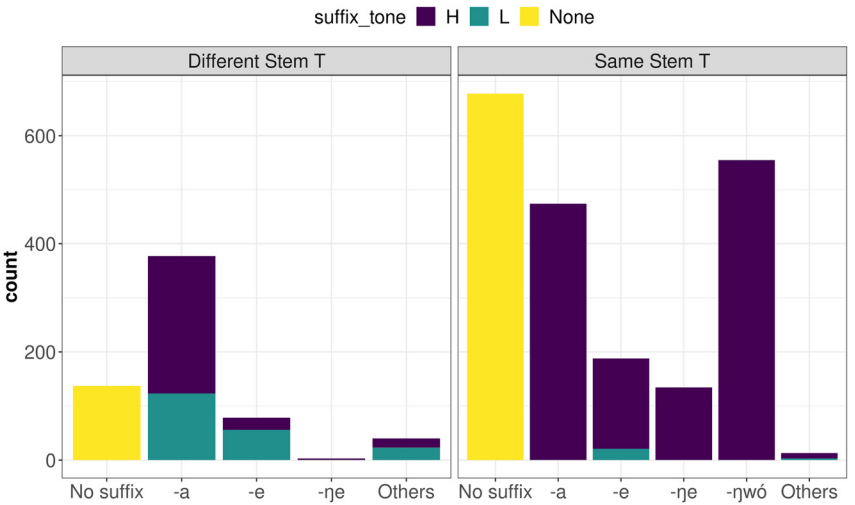

Figure 3: Changes observed between subject and object case forms.

occurs on the stem of the word, when inflected for object case. Finally, the eighth column mentions the change of tonal pattern on the stem if there is one. By way of illustration, in the fourth row, the original subject tonal scheme HH changes to LL in the object case. The cell of the eighth column is thus filled as 'HH-LL'.[4]

The segmental and tone changes are encoded in such a way as to reflect the information needed to derive object case paradigms. On the one hand, it is necessary to know (i) if a suffix is added and (ii) if so what its tone is. On the other hand, it is important to know (i) if the stem changes its tonal patterns and (ii) if so how. The length of the word is potentially relevant to infer the changes of tonal patterns. A glimpse at the distribution of these variables shows that segmental changes involve less complexity in comparison with tone changes. Indeed, a majority of segmental changes occur in word-final position. As shown in Figure 3, five main types of segmental changes are found: 'no changes' (i.e., the object case has exactly the same form as the subject case), 'adding a suffix /a/', 'adding a suffix /e/', 'adding a suffix / ŋwó/', and 'adding a suffix /ŋe/'. Only 2% (53/2,677) of the phonemic changes are either (i) expressed by means of another suffix or (ii) found word-internally. Due to their scarcity, the remaining suffixes and word-internal

4 The seventh and the eighth columns contain similar information; however the two columns were kept to assess if the presence/absence of tone change on the stem (column 'Stem T') is sufficiently helpful for predicting the tonal paradigms or if more details about the tone changes (column 'Stem T change' are required.

Table 2: Encoding of subject and object forms in Koalib. The abbreviation T refers to tones. The column 'Meaning' is added for the reader's comfort but is not included in the actual data.

Subject	Object	Meaning	Suffix	Suffix T	Stem syllable	Stem T	Stem T change
Kwéccè	Kwéccèŋwó	third-born girl	-ŋwó	H	2	Same	None
lɔ̀bbɔ̂l	lɔ̀bbɔ̂lá	trap	-a	H	2	Same	None
pèɽèár	pèɽèáré	ray (light)	-e	H	3	Same	None
káɗɗény	kàɗɗènyá	blind person	-a	H	3	Different	HH-LL

changes are marked as 'Others' in the data. Conversely, tone changes involve more complex patterns. The suffixes generally carry a tone that can vary depending on the noun. For instance, *báél* HH 'ghost, spirit' changes to *bàèlá* LLH but *léeɽé* HH 'rock' changes to *lèeɽéà* LHL. Both nouns have an additional suffix in the object case. However, the first noun has a suffix /a/ with a H tone while the second one has a suffix /a/ with a L tone. More than half (61%, 1,636/2,677) of the object forms take an extra H tone, the two other main categories being an extra L tone (8%, 226/2,677), and 'no additional tones' (31%, 815/2,677), i.e., zero suffix.

The main categories of changes from subject to object case are then identified based on the extracted information. Previous studies performed this task manually based on linguistic knowledge (Boychev 2013). In the current study, the categories are identified with automatic clustering. This choice is motivated by the fact that the data in this study is much larger and complex than previous studies and there is no existing study that provides a clear classification of nouns based on their different types of object case marking. First, a Gower distance matrix (Gower 1971) is generated based on the categorical variables listed in Table 2 (i.e. all columns except Subject, Object, and Meaning, which are raw data). Then, this distance matrix is fed to a hierarchical clustering algorithm. The algorithm starts by assigning each token to its own cluster. Then, the closest pairs of points based on the distances from the distance matrix are merged, and as a result the number of clusters goes down by 1. The distance between the new clusters is then computed to update the distance matrix. These two steps are repeated until all the tokens in the data are merged into one cluster. During this process, the silhouette method (Kaufman and Rousseuw 1990) is used to pinpoint the ideal amount of clusters. In our case, the ideal amount of clusters is five since the silhouette width reaches plateau at the number of five clusters. A detailed view of the generated clusters is displayed in Table 3. The first cluster is further divided into subclusters based on their similarities. It regroups nouns that tend to fulfill the following two conditions: (i) their stem does not undergo any change of tonal pattern and (ii) they either have no object suffix or take an object suffix with a H tone. While the silhouette method

Table 3: An overview of the clusters for case paradigms in Koalib. The abbreviation T refers to tone. Due to space limitation, only the largest categories of each variable are listed (for more details about the content of each cluster, see Section 4.1).

Cluster	Suffix	Suffix T	Stem T	Stem T change	Size	Ratio
1	- (100%)	None (100%)	Same (100%)	None (100%)	678	0.253
1a	-a (100%)	H (100%)	Same (100%)	None (100%)	474	0.177
1e	-e (100%)	H (90%)	Same (100%)	None (100%)	205	0.070
1ŋwó	-ŋwó (100%)	H (100%)	Same (100%)	None (100%)	555	0.207
1ŋe	-ŋe (100%)	H (100%)	Same (100%)	None (100%)	136	0.050
1Others	Others (100%)	H (10 items)	Same (100%)	None (100%)	13	0.005
2	-a (62%), -e (28%)	L (100%)	Different (100%)	HL-LH (59%), LL-LH (16%)	200	0.075
3	- (100%)	None (100%)	Different (100%)	HL-LL (59%), LL-HL (25%)	137	0.051
4	-a (98%)	H (100%)	Different (100%)	HH-LL (46%), LL-HL(15%)	260	0.097
5	Others (100%)	H (89%)	Different (100%)	HL-LL (8 items), F-L (5 items)	19	0.007

suggests keeping this cluster as one unit, we follow the additional splits from hierarchical clustering and split this cluster into subclusters during the analysis to increase the transparency of the linguistic analysis. That is to say, from a linguistic point of view, it is important to differentiate the cases where the stem does not undergo any tone change but takes different suffixes. Therefore, the first cluster is thus further split into subclusters that represent the use of different suffixes.[5] Cluster 2 regroups nouns that tend to fulfill the following three conditions: (i) their object suffix is either /a/ or /e/, (ii) their object suffix carries a L tone and (iii) their stem undergoes a change of tonal pattern. Cluster 3 has nouns that strictly fulfill the following two conditions: (i) they do not take suffixes in the object case but (ii) their stem undergo a change of tonal pattern when inflected for object case. Cluster 4 comprises nouns that strictly fulfill the following two conditions: (i) they take a /a/ suffix with a H tone and (ii) their stem undergo a change of tonal pattern when inflected for object case. Finally, cluster 5 regroups the 'outliers' that display word-internal changes and/or exceptional suffixes for the object case.

To sum up, the first cluster and its subclusters represent the nouns that are the most regular. These nouns do not change the tonal pattern of their stem and either

5 In the following text, we use the term 'cluster 1' (or alternately 'subcluster 1') for referring to the first of these subclusters, i.e., the cluster in which there is no suffix and no tone change on the stem. We use the term 'first cluster' for referring to the whole group of subclusters whose name begin with the digit 1, i.e., 1 + 1a + 1e + 1ŋwó + 1ŋe + 1Others.

(i) they take a suffix or (ii) they do not have any suffix in the object case, e.g. *ŋèráaɽà* 'sauce', which has only one form coding for both subject and object cases. Cluster 5 represents the few outliers of the system that have word-internal changes and/or exceptional object suffixes. Nouns that do not take suffixes but have a change of tonal pattern are regrouped in cluster 3, e.g., the subject noun *kwìcì* 'human being' and its object counterpart *kwícì*. Clusters 2 and 4 relate to nouns that mostly take the /a/ suffix and have a change of tonal pattern on the stem. The tonal changes can also be roughly identified per cluster. For the tonal patterns HL and LHL on the subject case, nouns in cluster 2 change to LH and LLH respectively for their object case while nouns in cluster 3 change to LL and LLL for their object case (see also Section 4.1 below). Cluster 4 mostly involves nouns that have consecutive H or L tones on their subject case and take the /a/ suffix with a H tone, e.g., the subject noun *báél* 'ghost, spirit' and its object counterpart *bàèlá*. An in-depth linguistic analysis of these clusters is provided in Section 4.

The main task of the classifier is to identify these clusters of case paradigms based on information from the non-inflected (subject) nominal root, using factors considered to have a significant impact on case-object inflection within the scope of previous studies (Boychev 2013; Quint 2010b). The following information is extracted automatically from the subject forms: word length, syllable structure, tone structure, final tone, final phoneme tone, and final phoneme. A sample of the final coding is shown in Table 4.[6] Word length refers to the total number of phonemes (or skeletal positions) in a noun. For instance, the word length of *Áacè* is 4. Syllable structure refers to the skeletal annotation of nouns, where C stands for any consonant and V for any vowel. As an example, the syllable structure of *Áacè* is VVCV. Tone structure refers to the entire tonal pattern of the noun, e.g., the tone structure of *Áacè* is HL. Final tone refers to the last tone of the noun. As an example, the final tone of *Áacè* is L. An additional variable is added to indicate if the final phoneme of a noun carries a tone or not. For instance, the noun *báél* ends in the consonant /l/, which in Koalib is not associated with tone. The last phoneme of each noun is also annotated without the tone it may bear. As an example, the last phoneme of *Áacè* is annotated as /e/. While the variable about the tone carried by the final phoneme is likely to share information with the variable of final phoneme (the final phoneme is more likely to carry a tone if it is a vowel), the two variables are still both considered since some words may have a final vowel that does not carry a tone in the transcription system designed for the corpus. For example, *Kéelàe* 'name of a neighborhood'. Finally, a given Koalib noun is considered as 'saturated' if it contains a VCCV or VVCV sequence in its stem. This factor is added since saturation seems to be an important criterion of well-formedness for

6 The format of the data has been slightly modified to fit in the table, the original format of the data displays items as rows and features as columns.

Table 4: Encoding of subject forms in Koalib. The abbreviations are read as follows: pr = proper noun, cn = common noun, A = animate, IN = inanimate, NM = not mentioned.

Type	Example 1	Example 2	Example 3
Subject form	Áacè	àatòm	léóm
Meaning	Aisha (name)	termitary	bird sp.
Etymology	Arabic	Koalib	Koalib
Noun type	pr	cn	cn
Noun class	kw_0	w_y	l_ŋw
Animacy	NM	IN	A
Word length	4	5	4
Syllable structure	VVCV	VVCVC	CVVC
Tone structure	HL	LL	HH
Final tone	L	L	H
Final phoneme tone	L	None	None
Final phoneme	e	m	m
Saturated	Yes	Yes	No

nominal, adjectival and verbal items in Koalib. The information on etymology, noun type, noun class, and animacy is also extracted from the data.

It is important to point out that the information used to generate the clusters (Table 2) involves the use of different suffixes, as well as tone change between subject and object case, among others. This information involves details about the object case itself. However, the information fed to the classifier (Table 4) is restricted to information on the subject case, e.g., word length, CV structure, tone structure, among others. That is to say, we are not using the same information to generate the clusters and to identify them later on. For example, the information about noun type is not used for generating the clusters but is fed to the classifiers.

As a summary, the different object case paradigms are extracted automatically by comparing the subject cases with the object cases. The main case paradigms are identified with clustering methods. The classifier is then asked to predict the cluster affiliation of each noun. The following subsection provides an overview of the classifier used in this study.

2.2 Overview of the method

To allow for the extraction of transparent rules that could be assessed from a linguistic perspective, deep learning approaches (Aharoni and Goldberg 2017; Cotterell et al. 2018; Kann and Schutze 2016; Makarov and Clematide 2018) are not used in the current analysis. On the other hand, we use classification methods that

are more transparent as to how the decisions are made by the model (Ahlberg et al. 2015; Sorokin 2016). More precisely, two decision-tree based classifiers are selected to identify the interaction of the variables and their relative importance within the data set. The first classifier is a single decision tree while the second classifier is composed of several trees. For the first classifier, the decision tree is generated with binary recursive partitioning (Breiman et al. 1984). That is to say, the data is consistently partitioned to form binary groups that are as homogeneous as possible. During each partitioning step, each variable is assessed and the variable that can result in the most homogeneous split is used. This process is repeated until the data cannot be split further.[7] The decision tree is expected to show the hierarchical interaction of the variables within the dataset. For instance, if both the etymology and the animacy have a significant effect on distinguishing clusters, the decision tree will show which of the two features has a prior predictive effect when they are both considered.

The second classifier functions in a similar way as the first classifier but builds a sample of 500 trees instead of only one tree, hence its name of *random forests*.[8] For each tree, a bootstrap sample of the entire data is used and a sample of the variables is selected. That is to say, each tree in the sample of 500 trees is built with a different bootstrap sample taken from the original data. For each tree, about one-third of the cases are left out of the bootstrap sample and not used in the construction of the tree. Then, the left out cases are used as a test set to assess the performance of the tree and calculate the accuracy/error rate of the predictions. There is thus no need for cross-validation or a separate test set to get an unbiased estimate of the test set error, since it is estimated internally during the run. This process of random sampling represents the main strength of decision-tree based classifiers, as it makes them applicable on small-scale data and takes into account the possible auto-correlation between variables (Tagliamonte and Baayen 2012). The random forests classifier is also used to assess the relevance ranking of the variables, which is obtained by calculating the average difference between the estimate and the out-of-bag error without permutation. The larger the importance of a variable, the more predictive it is. By way of illustration, if the accuracy of the

7 The hyper-parameters of the decision tree were set to its default as defined in the rpart R package (Therneau and Atkinson 2019). The minimum number of observations that must exist in a node in order for a split to be attempted was set to 20. The minimum number of observations in any final node was set to the rounded value of 20/3 = 7. The maximum depth of the decision tree was set to 30 nodes.

8 The hyper-parameters of the classifier were set to its default, as defined in the randomForest R package (Liaw and Wiener 2002). The number of trees was set to 500. The number of variables randomly sampled as candidates at each split was set as the square root of the number of variables in the data.

classifier drops the most when it does not take into account the etymology, etymology is considered to have the highest ranking within all of the variables.

The performance of the classifiers is assessed with two measures, the *f*-score and the accuracy. On the one hand, the accuracy indicates the performance of the classifier on the entire dataset. It is equal to the ratio of all the correctly retrieved tokens within the entire data. This value is compared with two baselines: the chance baseline and the majority baseline. The chance baseline represents the accuracy a classifier would get by doing random guesses. This can be calculated as the probability of the model predicting each cluster value multiplied by the probability of observing each cluster occurrence. Based on the size of the clusters listed in Table 3, this gives $0.253 \times 0.253 + 0.177 \times 0.177 + 0.07 \times 0.07 + 0.207 \times 0.207 + 0.05 \times 0.05 + 0.005 \times 0.005 + 0.075 \times 0.075 + 0.051 \times 0.051 + 0.097 \times 0.097 + 0.007 \times 0.007 = 16.5\%$. The majority baseline relates to the biggest category in the dataset. Since the biggest cluster is cluster 1 (25.3%, 678/2,677), the computational classifier could reach an accuracy of 25.3% simply by labeling all the nouns as belonging to cluster 1. Thus, the accuracy of a classifier should exceed 16.5% (the random baseline) to be considered as acceptable and exceed 25.3% to be considered as having good discriminatory power. On the other hand, the *f*-score (Ting 2010) is a combination of two other measures: precision and recall. Precision is the percentage of correct predictions of a targeted category out of all predictions of that category, whereas recall quantifies how many tokens are correctly retrieved among all the expected correct output. The two measures evaluate the output from two different perspectives. These two measures are then combined into the *f*-score, which is the harmonic mean of the precision and recall, i.e., 2(recall × precision)/(recall + precision).

The quantitative analyses of this paper were conducted with the following R (R-Core-Team 2021) packages listed in alphabetical order: cluster (Maechler et al. 2019), corpus (Perry 2017), data.table (Dowle and Srinivasan 2019), factoextra (Kassambara and Mundt 2020), parsnip (Kuhn and Vaughan 2019), random (Eddelbuettel 2017), randomForest (Liaw and Wiener 2002), randomForest explainer (Paluszynska and Biecek 2017), readr (Wickham et al. 2018), recipes (Kuhn and Wickham 2019), rpart (Therneau and Atkinson 2019), rpart.plot (Milborrow 2019), rsample (Kuhn et al. 2019), Rtsne (Krijthe 2018), stringr (Wickham 2019), tidyverse (Wickham 2017).

3 Results

Two main results are obtained via the classification task. On the one hand, the interaction of the variables in the entire dataset is visualized through a representative tree. On the other hand, the predictive power of all variables combined

and/or taken individually is extracted by using the random forests classifier. It is important to point out that the output of the single decision tree is mainly used as a visualization tool to display the process of generating decision trees. The random forest classifier is more powerful and is expected to capture more complex information that the first classifier could have missed. The output of the random forest classifier is thus the main component considered in our analyses.

3.1 Single decision tree

The model is first trained with 70% of the data to generate a decision tree. Then, this tree is tested on the remaining 30% of the data. In other words, the decision tree is used to predict the object case paradigms of the tokens in the test set, tokens which have never been seen by the model. During this process, the train and test sets are generated to maintain the ratio of each cluster as in the entire data set. For instance, since cluster 1 represents 25.3% of the data, the same ratio is found within both the train and test sets. The decision tree generated based on the training set is shown in Figure 4, which can be read as follows: if the branch goes toward 1 (bigger than 0.5) it means TRUE, if the branch goes toward 0 (smaller than 0.5), it means FALSE. For instance, starting from the root, if a noun is a proper noun (node 1 to 3), it is affiliated to cluster 1ŋwó. The accuracy of the prediction is shown in the final node (node 3). This prediction classifies 385 tokens, among which 384 are correctly classified.

In general, the tree indicates that information on the tone of the lexeme is the most relevant in identifying clusters, since most information listed in the tree is related to tones. For instance, the HH and LL patterns are generally affiliated to cluster 4. Information on etymology, word length, final phoneme, and noun type is also found in the tree. The other variables that are not shown in the tree are considered as not relevant by the classifier. A more detailed linguistic analysis of the decision tree is provided in Section 4.

To be sure that this tree is reliable, we need to assess its performance on predicting the cluster affiliation of nouns. The accuracy of the decision tree on this task is 78.2%, which is far above the random baseline (16.5%) and the majority baseline (25.3%).[9] This shows that the classifier can actually find information relevant for identifying object case paradigms in Koalib. For the current sampling, the training and the test sets comprise of 1,876 and 801 tokens respectively. The confusion matrix

9 In order to avoid the risk of biasing the results by using any specific training and test sets, 10 different pairs of training and test sets were generated by randomly sampling the whole data set and used to evaluate the classifiers. Due to the stability of the results across the 10 sampling processes (average accuracy 79% with standard deviation of 0.8%), we can infer that the output of the classifier is reliable. The results reported in this subsection are from the tenth sampling.

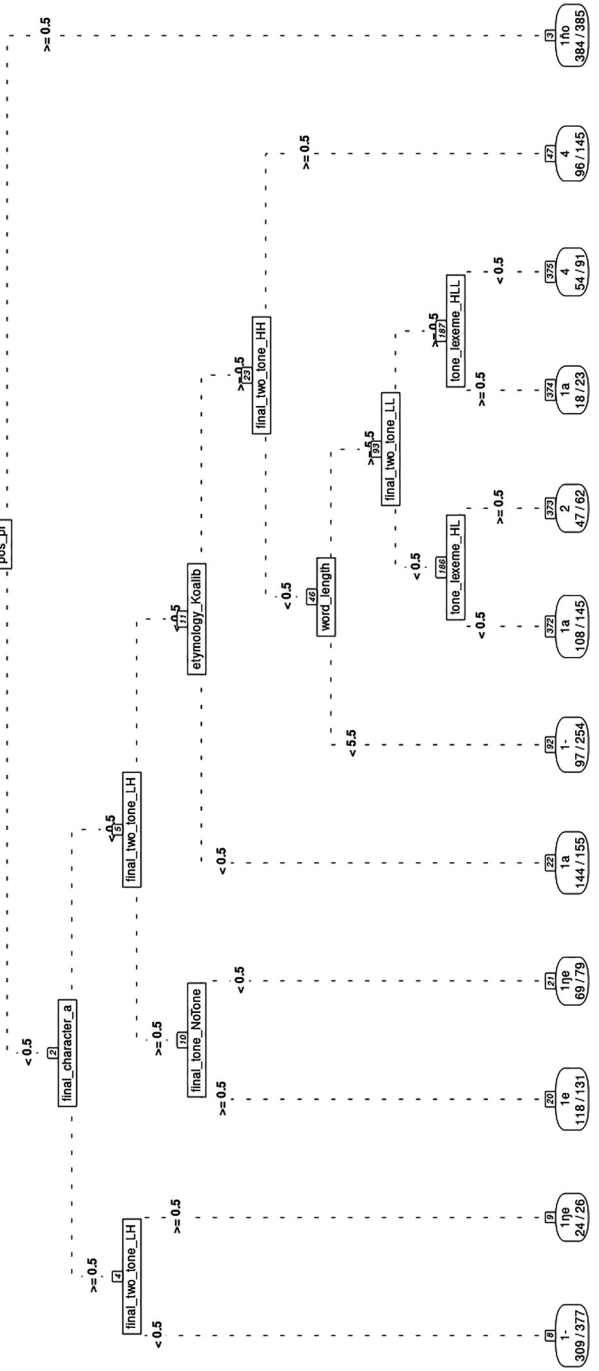

Figure 4: The decision tree of object case paradigms in Koalib. This tree is generated on the training set, which equals to 70% of the entire data. The abbreviations 'pr' refers to 'proper noun'.

Table 5: The confusion matrix of the decision tree on the test set. The columns are the predicted values and the rows are the actual values. The last three columns show the precision (Pre), recall (Rec), and *f*-score (F) on each cluster.

Cluster	1	1a	1e	1ŋwó	1ŋe	1Oth	2	3	4	5	Pre	Rec	F
1	178	5	1	1	2	0	1	0	21	0	0.650	0.852	0.737
1a	17	98	3	0	0	0	21	0	6	0	0.916	0.676	0.778
1e	5	0	48	0	0	0	0	0	1	0	0.923	0.889	0.906
1ŋwó	1	0	0	162	0	0	0	0	0	0	0.994	0.994	0.994
1ŋe	3	0	0	0	37	0	0	0	0	0	0.860	1.000	0.925
1Oth	2	1	0	0	1	0	0	0	1	0	0.000	0.000	0.000
2	16	1	0	0	0	0	42	0	7	0	0.656	0.636	0.646
3	40	2	0	0	0	0	0	0	1	0	0.000	0.000	0.000
4	8	0	0	0	3	0	0	0	61	0	0.622	0.847	0.717
5	7	0	0	0	0	0	0	0	0	0	0.000	0.000	0.000

shown in Table 5 matches with the decision tree and shows that the single tree classifier is good at discriminating between tokens from clusters 1 and its subclusters, 2, and 4. However, it seems to be pretty bad at identifying tokens from clusters 3 and 5, as no correct tokens are identified for these two clusters.

The performance is good for most of the clusters that do not involve tone change on the stem (e.g., cluster 1, 1a, 1e, among others). This is not surprising since these clusters involve less changes for the object. The acceptable performance on clusters 2 and 4 is also expected since these two clusters have different behaviors in terms of the suffix and tone they take. Clusters 3 and 5 are basically merged with cluster 1 by the classifier. The bad results on these two clusters can partially be explained by their size, as they are the smallest clusters in the data and thus provide less training material for the classifier. In terms of precision, recall, and *f*-score, the precision is higher than the recall for cluster 2, while the recall is higher than the precision for cluster 1 and 4. This observation shows that the model is generally correct when guessing that a noun belongs to cluster 2. However, it does not find all nouns from this cluster. As an example, the classifier guesses that 64 nouns belong to cluster 2 (column 2). Among these 64 nouns, 66% (42/64) are actually part of cluster 2. However, within all 66 nouns actually belonging to cluster 2 (row 2), only 64% (42/66) are found by the classifier.

3.2 Random forests

The second classifier provides information about how relevant each variable is for the classification task. The classification accuracy of the model is 83.3%, which is

Table 6: The confusion matrix of the random forests on the entire data set. The columns are the predicted values and the rows are the actual values. The last three columns show the precision (Pre), recall (Rec), and *f*-score (F) on each cluster.

Cluster	1	1a	1e	1ŋwó	1ŋe	1Oth	2	3	4	5	Pre	Rec	F
1	592	16	2	2	7	0	9	12	38	0	0.746	0.873	0.805
1a	16	396	9	1	0	0	40	0	12	0	0.886	0.835	0.860
1e	15	1	177	0	0	0	2	0	10	0	0.917	0.863	0.889
1ŋwó	3	1	0	551	0	0	0	0	0	0	0.993	0.993	0.993
1ŋe	0	0	1	0	135	0	0	0	0	0	0.900	0.993	0.944
1Oth	5	4	1	0	2	0	0	0	1	0	0.000	0.000	0.000
2	38	12	0	0	0	0	136	0	13	1	0.687	0.680	0.683
3	95	6	0	1	1	0	3	29	2	0	0.707	0.212	0.326
4	23	10	3	0	5	0	3	0	215	0	0.724	0.827	0.772
5	7	1	0	0	0	0	5	0	6	0	0.000	0.000	0.000

slightly higher than the accuracy of the single decision tree. This implies that using a more complex tool does capture additional information from the data, but not much. Additional variables probably need to be added so as to significantly increase the performance of the model. The confusion matrix and the precision/recall/*f*-score for the predictions of the random forests is shown in Table 6.

In general, we observe a performance similar to the single decision tree. Both classifiers can identify nouns from the first cluster and its subclusters, alongside with nouns from cluster 2 and 4. However, they fail to distinguish nouns from clusters 3 and 5. Nouns from both clusters 3 and 5 are almost entirely wrongfully assigned to cluster 1. Nonetheless, we also observe an improvement of performance for clusters 3 and 4, as their precision augments in comparison with the single decision tree.

We can also visualize how 'certain' the model is when making decisions (Figure 5). With random forests, the certainty level of the decisions is extracted by the probability of votes across all trees. As an example, if 400 of the trees assign a token to cluster 1, then the certainty of the decision is 400/500 = 80%. We see that the model generally has a certainty level over 60% for correct decisions (except for cluster 3) and has a certainty level ranging between 30 and 0% when wrong decisions are made. This indicates that the model is almost certain when making correct predictions and in doubt when making wrong decisions. This distribution reflects that the model is going in the right direction: it is more certain about decisions that turn out to be correct while it also knows that the decision is likely to be wrong when it actually makes wrong guesses.

The individual importance of the variables can be assessed via the conditional permutation-based variable importance. This process is expected to provide a

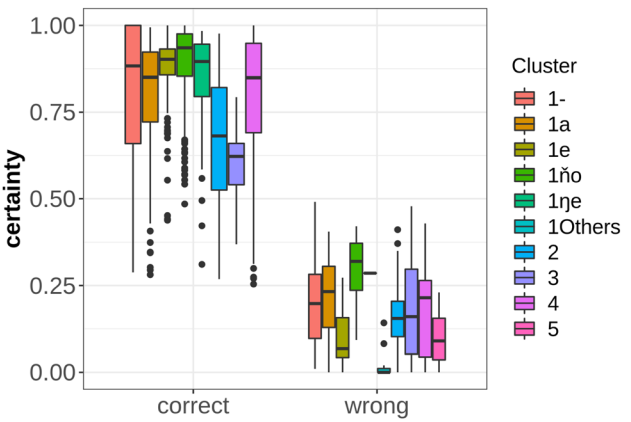

Figure 5: The certainty of the predictions from the random forests classifier.

more faithful representation of the predictive power of the variables. If a variable is consistently helpful in predicting the case paradigm of nouns in most of the data subsets, it implies that this variable has a high importance for the classification task. First, the frequency and the mean of the minimal depth for each variable within all the 500 trees generated by the random forests are visualized. The minimal depth indicates how far is the node with a specific variable from the root node, which is equal to a minimal depth of zero. If a variable is frequently close to the root node, it is considered to have a high importance. The minimal depth of the top 10 most important variables is shown in Figure 6. The main relevant variables are noun type, noun class, final phoneme, word length, and animacy. Noun types refer in particular to the distinction between proper nouns and common nouns. The noun class considered relevant by the model is `kw(sg)_0(pl)`. One final phoneme is pointed out by the model: /a/.

Partially similar results are found when using other measures. In Figure 7, the variables are ranked according to their effect on the accuracy and the purity of the nodes. On the one hand, the mean decrease of accuracy indicates how much worse the model performs without each variable. A high decrease implies that the variable has a strong predictive power. On the other hand, the mean decrease of the Gini coefficient indicates how each variable contributes to the homogeneity of the nodes and the end of the tree. A high decrease of Gini coefficient when removing a variable implies that this variable has a strong predictive power and therefore a high importance.

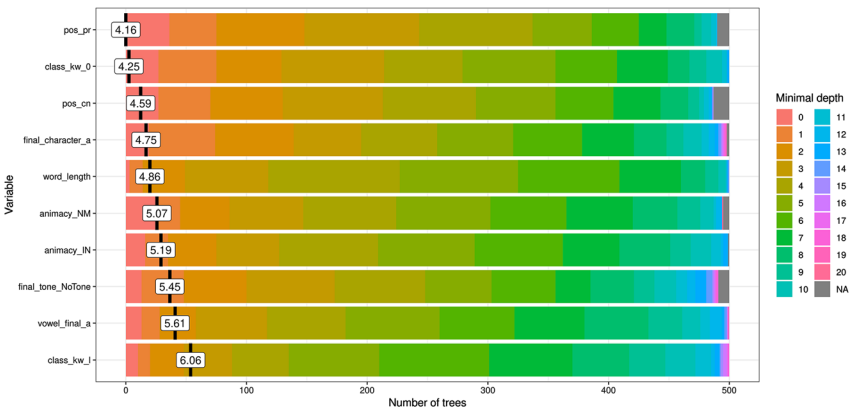

Figure 6: Distribution of minimal depth and its mean from the random forests classifier.

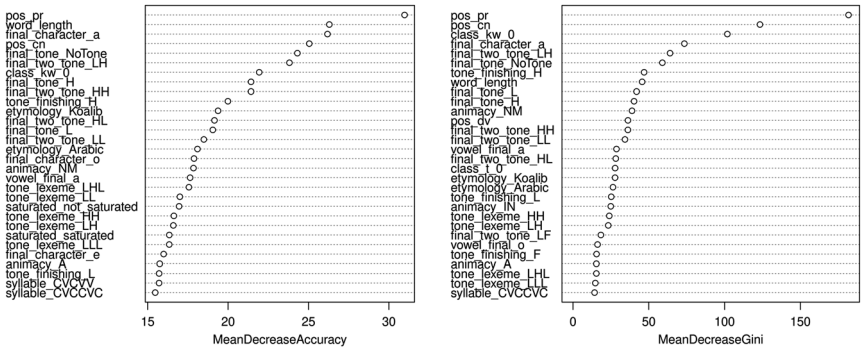

Figure 7: Accuracy and purity of the nodes from the random forests classifier.

One first observation is that the slope of accuracy, and Gini coefficient are rather different from minimal depth. On the one hand, the minimal depth shows a gradual decline. On the other hand, the decrease of Gini coefficient and accuracy is quite abrupt within the first six variables, while the other variables are almost on a similar level. In both three measures, the top six variables are the most relevant. As an example, the decrease of Gini coefficient is very steep until the sixth variable, after which it stabilizes. This means that removing the variables after the top sixth variable only results in a small loss of accuracy. In terms of variables, as shown in Table 7, the following items are consistently found within the top six variables of the three measures: Noun types (common vs. proper nouns), and the final phoneme /a/. This shows that these variables are considered to be the most stable contributors to the performance of the model.

To sum up, the two decision-tree-based classifiers have a similar performance of 78.2 and 83.3% on identifying the 10 clusters of object case paradigms in Koalib.

Table 7: The importance of the variables in the random forests classifier. The highlights in italics are the variables found in all three measures.

	Minimal depth	**Decrease accuracy**	**Gini coefficient**
1	*proper noun*	*proper noun*	*proper noun*
2	class kw_0	word length	*common noun*
3	*common noun*	*final phoneme a*	class kw_0
4	*final phoneme a*	*common noun*	*final phoneme a*
5	word length	final phoneme no tone	final phoneme tone LH
6	animacy unknown	final phoneme tone LH	final phoneme no tone

The single decision tree allows us to visualize the hierarchical interaction between the variables, while the random forests provide a ranking of importance for these variables. In the following section, we analyze these results from a linguistic perspective.

4 Linguistic interpretation of the results

In this section, we first provide a linguistic analysis of the clusters identified by hierarchical clustering. Then, we compare the results of the decision-tree based classifiers with previous studies. On the one hand, we compare the performance of our models with previous studies using rule-based classifiers. On the other hand, we compare the output of our models with linguistic studies to assess if our results match with the linguistic literature and can provide novel insight in identifying rules of object-case inflection in Koalib.

4.1 Linguistic analysis of the clusters

The clusters identified automatically (see Table 3) are extremely informative from a linguistic point of view. Of the 678 nouns deprived of any type of object inflection included in cluster 1, (i) 592 nouns (87%) end in a vowel, the overwhelming majority of which have a final low tone (514/592 = 87%). Furthermore, 450 of the 592 nouns (i.e. 76%) ending with a vowel end more specifically with a final /à/, e.g. *kérnà* (S = O) 'skin'. We can therefore infer that a final /a/ bearing a low tone tends to block the appearance of any object marker. If we consider (see Figure 3) that /a/ is the most common object suffix in the whole data, we can assume that a final /a/ blocks the appearance of an object suffix due to phonological rules (lack of contrast between the final phoneme of the subject form and the most common object suffix); (ii) regarding the minority of cluster 1 nouns ending with a consonant, a majority of them have a $(L)_n$ (with $n \geq 1$) tone profile and/or are unsaturated, i.e. they do not contain a VCCV or VVCV sequence in their stem (see above additional explanation before Table 4), e.g. *kèpèr* CVCVC (unsaturated) + LL (S = O) 'side', *lòkwɔ̀m* CVCVC (unsaturated)[10] + LL (S = O) 'crane sp. (bird)', *kéjáŋ* CVCVC (unsaturated) + HH (S = O) 'hot weather', *kàaràm* (saturated) + LL (S = O) 'tick (arachnid)'. For this subcluster, syllabic structure

10 /kw/ counts as one consonant in the Koalib phonological system (Quint and Ali Karmal Kokko 2009, pp. 108–109).

and tone pattern seem to be the main factors blocking the emergence of a case inflection.

Regarding the 474 nouns of cluster 1a, the case suffix is always a high-toned /a/. More than 86% (410/474) of the nouns have a final HL or F (ultimately analyzable as HL) tone, 306 nouns (i.e. 306/474 = 65%) end with a consonant, and 217 (i.e. 217/474 = 46%) have an exogenous origin (either Arabic and/or English) a proportion three times higher than the ratio of borrowings attested in the corpus (see above Section 2). In other words, cluster 1a seems to regroup two kinds of nouns: (i) items selected on a phonological basis, whose subject form ends with a consonant and/or whose subject tone pattern ends with HL or F, e.g. *lèrɔ́mpɔ̀l* (S) 'grass sp.' > *lèrɔ́mpɔ̀lá* (O), *kɔ̀ɔkɛ̂* (S) 'bird sp.' > *kɔ̀ɔkɛ̂á* (O) and (ii) items selected on the basis of their exogenous origin (borrowings), e.g. *céjèn*, realized as [ʃíɟìn] (S) 'jail' (< Sudanese Arabic *síjin*) > *céjènà* [ʃíɟìné] (O), *kɛ̂k* (S) 'cake' (< English 'cake') > *kɛ̂ká* (O).[11] Note that cluster 1a does not enclose any noun whose subject form ends with /a/, which confirms the observation made for cluster 1.

All 205 nouns of cluster 1e end with a consonant. They can be divided into two sets or subclusters according to the tone of their case suffix: (i) a majority (184/205 = 90%) have a high-toned /e/ suffix. Almost all of these nouns (176/184 = 96%) have a subject form ending with a LH sequence, e.g. *kɔ́llòkwɽɔ́r* HLH (S) 'wild bean' > *kɔ́llòkwɽɔ́ré* (O), and almost one half (81/176 = 46%) end with /ŋ/, e.g. *kwàaɽáŋ* LH (S) 'whea' > *kwàaɽáŋé* (O); (ii) the remaining 21 nouns have a low-toned /e/ suffix. All of them have a subject form that ends with a low tone (L) and all but two have low isotonic tonal patterns (i.e. L_n with $n \geq 1$), e.g. *lòmòr* (S) 'stick' > *lòmòrè* (O). Also, most of the members of this subcluster (13/21, i.e. two thirds) have unsaturated subject forms. In other words, both subclusters 1é and 1è are obviously quite consistent and their members seem to be ascribed to each of them on phonological bases.

As said above, cluster 1ŋwó obviously owes its existence to semantics, as nearly all of its 555 members (544/555, i.e. 98%) are proper nouns, mainly anthroponyms, e.g. *Kwókkò* (S) 'first-born male' > *Kwókkòŋwó* (O) and place names, e.g. *Kálkè* (S) 'Delami (Koalib city)' > *Kálkèŋwó* (S). As to the remaining members of the cluster, they can be accounted for by the fact that the object suffix /ŋwó/ is also used as a default marker to inflect non-prototypical nouns for object. In our data base, this applies to deverbal nouns such as *óŋnè* [úŋnì] (S) 'blackness' (< *óŋnè* [úŋnì] 'be(come) black') < *óŋnèŋwó* [úŋnìŋwú] (O) which, contrary to most deverbals, are produced directly from the centrifugal imperfective form of the verb

11 Note the fact that many exogenous nouns are found in cluster 1a is also linked with the high proportion of final HL and F tone sequences among these nouns. For more details about tonal integration and case inflection of Koalib borrowings, see Quint (2018).

without any added segmental morphology (i.e. by a mere conversion process). The only common noun (or non-deverbal) of the database, *nyɔ̀kkɔ̂ŋ* 'asking small amounts (of money or other things) to other people' is itself an expressive derivation from the adjectival basis *ɔ̀kkɔ̀* 'few' and therefore cannot be held either as a prototypical noun, which most certainly accounts for its being inflected for object with a /ŋwó/ suffix. Only two nouns labeled as proper nouns in the database do not take /ŋwó/ for their object case, namely *Állà* (S) 'God' (> *Àllà* (O), with tonal inflection) and *kícɛ̀* 'Fido (dog's name)'. Regarding *Állà* (borrowed from Arabic), the absence of /ŋwó/ is probably linked with the specific limits of the semantic category of proper nouns in Koalib, as the name of one of the most famous traditional deities of the Koalib country (located in the town of Dere (Arabic)/ *Kwánʈàŋ* (Koalib)), *tîiɽù* (S) (> *tîiɽùé* (O)) also has no /ŋwó/ object suffix. Therefore, what appears is that divine entities do not belong to the semantic category of proper nouns in Koalib and that *Állà* should probably be considered as a common noun in Koalib. The exceptional behaviour of *kícɛ̀* is harder to explain, as other frequent dog's names such as *Jákcɔ̀n* (S) (< *Michael Jackson*) (> *Jákcɔ̀nŋwó* (O)) do have a /ŋwó/ object.

Cluster 1ŋe is a very consistent category: all of its 136 members have a high-toned object suffix /ŋé/ and their subject form ends both with a vowel and with a LH or FH (ultimately analyzable as HLH) tone sequence, e.g. *kwèpàanyá* LLH (S) 'foreigner' > *kwèpàanyáŋé* (O), *kwôrró* FH (S) 'turtledove' > *kwôrróŋé* (O). Therefore, cluster 1ŋe is clearly linked to phonological factors.

Cluster 1Others regroups only a dozen nouns, which are clearly exceptions and do not seem to pattern together. Most of 1Others members display either an internal change of the stem, e.g. *lèr* (S) 'type of hat' > *lèrrè* (O), *kél* (S) 'seed hole' > *kéelé* (O) or an exceptional object suffix, e.g. *kwɛ́kkɛ̀* (S) 'member of a specific Koalib subtribe' > *kwɛ́kkɛ̀ŋá* (O) (suffix *ŋá*) or both, e.g. *kwào* (S) 'woman' > *kwàeò* (O) (suffix *ò* + change of the subject stem). Note that most nouns belonging to this cluster are monosyllabic.

Cluster 2 regroups several morphological patterns and subclusters, most of which exhibit /e/ or /a/ as object suffixes: (i) the dominant one (83/200 = 42%) comprises nouns whose subject form ends with a HL or HLL sequence, which are nearly all saturated (67/83), and which all end with a consonant. In this case, the object form always has a low-toned vowel and ends with a LHL sequence (for subject ending in HL) or LLHL sequence (for subject ending in HLL). From a tonal point of view, it seems that the tonal sequence HL just moves to the rightmost edge of the noun when inflected for object case. From a segmental point of view, if the last vowel of the subject form is /a/, then the object suffix is generally /e/, e.g. *káaŋàl* HL (S) > *kàaŋálè* LHL (O), *ʈàbéllàŋ* LHL (S) 'little finger/toe' > *ʈàbèlláŋè* LLHL (O). Conversely, if the last vowel of the subject form is not /a/, then the case

suffix is generally /a/, e.g. *lɛ́tmɛ̀n* HL (S) 'bean' > *lɛ̀tmɛ́nà* LHL (O), *kwótkòròny* HLL (S) 'member of a specific Koalib subtribe' > *kwòtkòrónyà* LLHL (O); (ii) the second subcluster comprises with 38 nouns (38/200 = 19%) whose subject form has an isotonic LL or LLL tone-pattern, a generally unsaturated syllabic structure (25/38) and in most cases (32/38) a final consonant. From a tonal point of view, the object form is regularly (L)LHL, where the last syllable of the stem is raised to H and the object suffix bears a low tone. From a segmental point of view, if the last vowel of the subject form is /a/ or if its last consonant is /n/ or /r/ (i.e. a dental non-obstruent), the case suffix is generally /e/, e.g. *kèbàŋ* LL (S) 'cave' > *kèbáŋè* LHL (O), *ʈòɽòn* LL (S) 'ground squirrel' > *ʈòɽónè* LHL (O), *kwòɗòr* [kwùɗùr] LL (S) 'monster' > *kwòɗórè* [kwùɗúrì] LHL (O). Conversely, if the last vowel of the subject form is not /a/ and if the last consonant is not /n/ or /r/, then the object suffix is always /a/, e.g. *lèrɔ̀ny* LL (S) 'spring (water)' > *lèrɔ́nyà* LHL (O); (iii) still another clear subcluster appears for 8 nouns (4% of cluster 2), whose subject form has a HL tone pattern, an unsaturated syllable structure and a final consonant and whose object form has a /e/ suffix and a LLL tone pattern, e.g. *kóròn* HL (S) 'wind' > *kòrònè* LLL (O).

Cluster 3 regroups the 137 items whose case inflection is expressed only by a tone change. A huge majority of these items end with a vowel (134/137 = 98%), mostly /a/ (94/137 = 69%), and they are saturated (116/137 = 85%). Two main subclusters can be distinguished: (i) a dominant one (83/137 = 61%) whose subject tone contains a HL or D sequence that is lowered to LL or L in the object form and where final /a/ is clearly dominant (75/83 = 90%), e.g. *kéŋlà* HL (S) 'sorghum spikelet' > *kèŋlà* LL (O), *kèpérttà* LHL (S) 'river bed' > *kèpèrttà* LLL (O), *kwóntònà* HLL (S) 'member of a Koalib tribe' > *kwòntònà* LLL (O), *têa* D (S) 'tail' > *tèa* L (O), *kâaɽà* DL (S) 'splinter' > *kàaɽà* LL (O); (ii) a subcluster (38/137 = 27%) characterized by the opposite tonal profile, where the subject form contains a LL or L sequence which changes to HL or F (i.e. an underlying HL) respectively and where final /a/ is not the norm (12/38 = 32%), and whose members are nearly all disyllabic (three exceptions), e.g. *kɛ̀lmɛ̀* LL (S) tortoise > *kɛ́lmɛ̀* HL (O), *ŋèa* L (S) 'poison' > *ŋêa* F (O).

Cluster 4 regroups almost exclusively (254/260 = 98%) items whose object forms have a final suffix /a/ with high tone. Several subclusters can be distinguished: (i) 145 items (i.e. 145/260 = 56%), nearly all saturated, have an isotonic subject stem $(H)_n$ (with $n \geq 2$) that becomes $(L)_n$ when inflected for object, e.g. *káŋkór* [kéŋgúr] HH (S) 'hyena sp.' > *kàŋkòrá* [kèŋgùré] LLH (O), *lébárttó* HHH (S) 'round pebble' > *lèbàrttòá* LLLH (O); (ii) 88 items (i.e. 88/260 = 34%), nearly all saturated, have an isotonic subject stem $(L)_n$ that becomes $H(L)_{n-1}$ when inflected for object, e.g. *lòrkò* LL (S) 'knee' > *lórkòá* HLH (O), *kèjɛ̀ɛɽɛ̀ny* LLL (S) 'stream' > *kéjɛ̀ɛɽɛ̀nyá* HLLH (O); (iii) 17 items (i.e. 7%) have a subject form with a HL profile, a generally unsaturated stem (only two exceptions), a final vowel

(one exception), and in most cases a final long vowel or a final vowel sequence of at least two vowels (four exceptions). When inflected for object, the stem takes on a LL tone pattern, e.g. *kwétɛ̀ɛ* HL (S) 'antelope sp.' > *kwètɛ̀ɛ́á* LLH (O), *kémào* [kímèu] HL (S) 'snake' > *kèmàoá* [kìmèué] LLH (O).

Cluster 5 includes 19 items: (i) 10 are monosyllabic, all of which display a stem change, e.g. *lân* (S) 'sorghum grain' > *làaná* (O) and most of which also have exceptional case suffixes, e.g. *kwór* (S) 'man' > *kwòoró* (O), *ŋên* [ŋîn] (S) 'blood' > *ŋèenáŋé* [ŋìinéŋí] (O); (ii) all remaining nouns are disyllabic and at least five are slightly irregular variants of subcluster (iii) of cluster 4, e.g. *kéɽòo* HL (S) > *kèɽòá* LLH (O) (expected **kèɽòoá* according to the main pattern of cluster 4, subcluster (iii)).

The different morphological patterns discussed across the clusters are summed up in Table 8. If we leave aside cluster 1Others and 5, which are catch-all categories, the other eight clusters (1, 1a, 1e, 1ŋwó, ŋe, 2, 3, 4) enclose 17 subclusters and dominant morphological patterns (such as (i) + à and (i) + è for cluster 2). Some of these clusters (e.g. 1ŋwó and 1ŋe) were already identified in previous study (Quint 2010b). However, the present clusterization has undoubtedly enriched and refined the analysis and contributed to better distinguish and individualize these morphological patterns. It also helps to grasp the most significant characteristics of the subject form which account for the actual form of the object of a given noun in Koalib. These characteristics belong to various linguistic levels, such as (i) phonology: lexical tone (and in particular final tone(s)); final phoneme(s); saturation; number of syllables; (ii) diachrony: endo- or exogenous origin; (iii) semantics and morphology: noun types (common vs. proper, derived vs. basic).

However, these many factors do not allow us yet to make accurate predictions in all cases: for example, the characteristics of the subject defined (see Table 8 above) for subcluster 3 (i), namely HL or D sequence + final /a/ are compatible with the characteristics of the subject defined for subcluster 1 (i), namely final –L and final /à/, hence the difficulty of the decision trees to recognize items belonging to cluster 3. In a similar vein, it is quite difficult to explain why a given item should belong to subcluster 1e (subject = $(L)_n$ + final –C + unsaturated) or to subclusters 2 (ii) +(è) or 2 (ii) +(à), whose characteristics for subject match those of subcluster 1e. Actually, what the clusters method was able to put in evidence is the existence of recurring subject-object pairings in Koalib and the fact that these pairings do cluster in quite consistent morphological patterns. Regarding its predictive power, the clusters method, despite the significant results that it has obtained, is still failing to predict the object form of a given Koalib item with a high degree of accurateness. These limitations may also be due to the power of the algorithm but also to the nature of the input given to the model (some new parameters might be

Table 8: The morphological patterns attested in the automatically identified clusters. The abbreviations are interpreted as follows: Sub = subcluster, INDIF = indifferent, NA = not attested, pr = proper noun, dv = deverbal, syl = syllable.

		Subject				**Object**	
Cluster	Sub	Tone	Final phoneme	Saturation	Other	Tone	Suffix
1	(i)	-L	a	INDIF	NA	= S	NA
	(ii)	$(L)_n$	C	-	NA	= S	NA
1a		-HL/-D	C and not /a/	INDIF	+exogenous	-HLH/-DH	á
1e	(i)	-LH	C (-ŋ)	INDIF	NA	LHH	é
	(ii)	$(L)_n$	C	-	NA	$(L)_{n+1}$	è
1ŋwó		INDIF	INDIF	INDIF	pr	S-H	ŋwó
1ŋe		-LH/-FH	V	INDIF	NA	-LHH/-FHH	ŋé
1Oth		INDIF	INDIF	INDIF	1 syl	INDIF	INDIF
2	(i) + à	$(L)_n$HL/ HLL	aC	+	NA	$(L)_n$LHL/ LLHL	è
	(i) + è	$(L)_n$HL/ HLL	VC and V ≠ /a/	+	NA	$(L)_n$LHL/ LLHL	à
	(ii) + è	$(L)_n$	aC Vn Vr	-	NA	$(L)_{n-1}$ HL	è
	(ii) + à	$(L)_n$	last V ≠ /a/, C ≠ /n,r/	-	NA	$(L)_{n-1}$ HL	à
	(iii)	HL	C	-	NA	LLL	è
3	(i)	.HL./.D.	a	INDIF	NA	.LL./.L.	NA
	(ii)	LL/ L	V/ VV	INDIF	2 syl/ 1 syl	HL/ D	NA
4	(i)	$(H)_n$	INDIF	+	NA	$(L)_n$H	á
	(ii)	$(L)_n$	INDIF	+	NA	H$(L)_{n-1}$ H	á
	(iii)	HL	VV	-	NA	LLH	á
5	(i)	INDIF	INDIF	-	1 syl	INDIF	INDIF
	(ii)	HL	VV	-	2 syl	LLH	á

taken into account) and/or to the sheer complexity of Koalib case morphology, which may display a fair amount of idiosyncrasies which simply do not fit the (known) rules.

4.2 Linguistic analysis of the decision trees

The high performance of the first cluster and its subclusters is linguistically expected, since it is more semantically transparent and/or more regular in terms of phoneme and tone changes. As an example, the nouns taking the suffix /ŋwó/ are mostly proper nouns (see above Section 4.1). Proper nouns are annotated within the noun type information, which provides a direct clue to the classifier. Moreover, the object case of proper nouns only involves the suffixation of /ŋwó/ and does not involve any other change in the lexeme. Thus, the classifier can very easily identify the nouns affiliated to the case paradigm of cluster 1ŋwó. The other clusters for which the classifier reaches an acceptable accuracy are also linguistically predicted clusters. As an example, cluster 2 and 4 include the object case paradigms mostly characterized by an /a/ suffix that carries either a L or a H tone and by a tone change on the lexeme. Cluster 3 is relatively harder to identify since its tokens are very similar to the tokens in cluster 1 (see final discussion in Section 4.1), the only difference being the change of tonal pattern on the stem for cluster 3. Finally, cluster 5 is extremely small and comprises most of the outliers of the data, which undergo a word-internal change in addition to suffixation. The relatively small size of cluster 3 and 5 is also likely to have a negative impact on the performance of the classifier.

If we consider the three top-recurring variables in the random forest classifiers (see Table 7), the distinction between proper versus common noun is clearly conspicuous from a linguistic point of view. It corresponds to the first node of the single decision tree (see Figure 4) and emphasizes the fact that virtually all proper nouns are inflected for object on a semantic basis and with a very recognizable suffix, namely /ŋwó/. The third variable, the presence (or absence) of a final phoneme /a/, is one of the main criteria used to define 3 of the 17 commonest morphological patterns for the object case in Koalib (see Table 8), i.e. subclusters 1 (i), 3 (i) (final /a/ dominant for both) and cluster 1a (final /a/ not allowed). Here again, the fact that final phoneme /a/ is analyzed by the random forests classifier as a significant variable and that at the same time it is a basic phonological criterion shared by both subclusters 1 (i) and 3 (i) explains why the classifier has such a low performance in identifying cluster 3 as a separate morphological category (see Tables 5 and 6).

4.3 Comparison with previous studies

In terms of accuracy, previous studies (Boychev 2013) obtained an accuracy of 64% on a data set of 1,200 nouns when only considering the contrast between H and L tones. Our experiments result in an accuracy of around 80% on a data set of 2,677 nouns while considering the difference between all four tones found in Koalib: H, L, F, and R. It is important to point out that our study used automatic clustering instead of manual labeling of tonal and segmental changes. The results are thus not entirely comparable, as Boychev (2013) worked on a database different from the one used here: in Boychev's database, (i) proper and exogenous (i.e. borrowed from Arabic and English) nouns had been excluded, (ii) a lexical root was counted only once even if it appeared in several lexical items, e.g. Koalib *kwór*, 'man', *ŋór* 'manhood' and *ʈór* 'child' are all derived from the same lexical root, *-ór* 'man' to which several class-prefixes are added (*kw-* for human beings, *ŋ-* for abstract nouns, *ʈ-* for diminutives) and the object case remains the same for all items derived from the same lexeme: *kwòoró* 'man' (O), *ŋòoró* 'manhood' (O), *ʈòoró* (O). In Boychev, *kwór*, *ŋór* and *ʈór* were counted as one item (because they share the same lexical root) whereas in the present study they were counted as three different items, as they appear as three different entries in the data and indeed are associated with different meanings, some of which are not necessarily transparent, e.g. *kwór* only refers to a 'male adult' while *ʈór* refers to any child, independently of their gender; (iii) the database used here is significantly more complete and reliable than the one used by Boychev, as Quint and Ali Karmal Kokko (2022) have since then been busy in increasing both the lexical coverage and the quality of the forms of case inflection for each nominal item. Be that as it may and in spite of these sampling differences, the present results admittedly show a significant improvement both from a quantitative and qualitative point of view.

If we compare our study with Quint (2010b), developed without the use of machine learning, the present study (as well as Boychev (2013)'s) has the main advantage of enabling a global coverage of all morphological patterns. Previous studies were able to identify several of the most common recurring morphological patterns (e.g. 1e (i), 1ŋwó, 1ŋe, 3 (i), 3 (ii), 4 (i) and 4 (ii) (see Table 8), but several others were missed due to the limits imposed on the human brain in capturing the many parameters or characteristics involved in the morphology of the object case in Koalib. This is particularly the case of the various morphological patterns and subclusters of cluster 2, which illustrate the role played by saturation and the last (non necessarily final) vowel of the subject form. The previous studies had always failed to describe so precisely the interplay of factors that characterize these subclusters. In terms of classification, on the one hand, our results corroborate

previous studies by showing that the change of tonal pattern on the stem is helpful for identifying object case paradigms (Boychev 2013, p. 41). For instance, clusters 1–4 are distinguished based on their different changes of tonal patterns on the stem. We also find that noun types are relevant for identifying case paradigms, as most proper nouns take the suffix /ŋwó/ and do not have a change of tonal pattern on the stem. Finally, we also find that shorter words (less than six phonemes or skeletal positions) are less likely to have tonal changes on the stem. On the other hand, in terms of novel contribution, the rules extracted from the decision tree cover a broader perspective than previous rules. Moreover, their interpretation through the decision tree allows us to avoid potential conflicts between rules, as the rules in the tree interact hierarchically and the linguistic analysis of the identified clusters has enabled us to distinguish many declensional morphological patterns that had not been previously described or identified.

5 Conclusion

We have shown that quantitative methods of clustering and classification can be used to identify morphological patterns and rules even in less-studied languages for which less (diverse) data might be available. While the results of these methods are supported by the output of our linguistic analysis, it is important to highlight that the core of the linguistic analysis is based on a solid theoretical underpinning and a thorough understanding of the language warranted by the use of a corpus, the native competence of our language consultants and a significant exposure of the first author to spoken Koalib and Koalib traditional life. The quantitative methods are only a supplementary tool that can help in generating and testing linguistic hypotheses. The results presented here clearly show that this cross-disciplinary approach is not only successful in verifying linguistic hypotheses, but also helpful in identifying new conditioning factors and interactions within language data.

In terms of limitation, as for any new methodology, there remains a number of important questions to clarify. In terms of linguistic analysis, a more refined synchronic and diachronic analysis of the identified clusters is required. In terms of quantitative analysis, additional methods and parameters should be considered. In the current study, we only explored the potential of one clustering method and two tree-based classifiers, without tuning their individual parameters. On the one hand, additional experiments should be conducted to evaluate the general performance of these methods. On the other hand, other methods of clustering and classification should also be tested to find the results with the best performance. The use of other methods of sampling and/or data treatment may also contribute to

the improvement of our results. The data used in this study also depended greatly on the intuitions of one main language consultant: the use of a wider sample of reference Koalib speakers should probably contribute to refining and developing our model in the future and check whether and how the present results are generalizable to the whole Koalib-speaking community.

6 List of abbreviations

CLF = noun class marker, DEM = demonstrative, PFV = perfective, PL = plural, PRF = perfect, PROX = proximal, SG = singular.

Acknowledgements: The authors are thankful for the comments of the editors and the reviewers, which helped to significantly improve the content of the paper. They are also grateful to Siddig Ali Karmal Koko for sharing with them his knowledge of and expertise on his mother tongue, Koalib.
Research funding: The first author is thankful for the support of the following grants: (i) PICS franco-soudanais *Les langues du Soudan: à la croisée des aires et types linguistiques* [The languages of the Sudan: a typological and areal crossroad]; (ii) PHC-Napata Kin terms and anthroponyms in the Nuba Mountain languages; (iii) Labex EFL, Strand 3, Workpackage RT1 – Language genealogy (Niger-Congo, Austronesian): Reconstruction, internal classification and grammatical description in the world's two biggest phyla: Niger-Congo and Austronesian (ANR-10-LABX-0083). This last grant contributes to the IdEx Université de Paris – ANR-18-IDEX-0001. The second author is also thankful for the support of grants from the Université de Lyon (ANR-10-LABX-0081, NSCO ED 476), the IDEXLYON Fellowship (2018–2021, 16-IDEX-0005), and the French National Research Agency (ANR-11-IDEX-0007, ANR-20-CE27-0021).

References

1967. *T̲ikitad̲iza t̲ian̲ [The New Testament in Koalib]*. Khartoum: The Bible Society of the Sudan.
1993. *Wa@d̲ wiyan̲ [The New Testament in Koalib]*. Khartoum: The Bible Society in Sudan.
Abdalla, Jummize & Abdalla Komi. 2000. *Yəwə na Nyaamin Nyəthi Kithilə Kir 2000 [A calendar for the year 2000, lit. 'Months and days of the year that is 2000']*. Khartoum: Khartoum Workshop Programme.
Abdalla Omer, Jummeiz, Abdalla Komi Kodi & Ibrahim El-Haimer. 1995. *Ŋwɔɔli Ŋwiyaŋ Kandisa-Gi Kəṯ̲hi Kouliib [A new Koalib alphabet]*. Khartoum: Kouliib Language Development Committee.

Abdalla Omer, Jummeiz, Shanan Suliman Kodi & Abdalla Komi Kodi. 1998. *Riŋerɔŋ Rəthi ŋwɔɔli ŋwiyaŋ kandisa-gi Kəthi kwəliib [A new Koalib alphabet illustrated by short stories]*. Khartoum: Kwəliib Language Development Committee.

Aharoni, Roee & Yoav Goldberg. 2017. Morphological inflection Generation with Hard Monotonic Attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2004–2015. Vancouver, Canada: Association for Computational Linguistics.

Ahlberg, Malin, Markus Forsberg & Mans Hulden. 2015. Paradigm classification in supervised learning of morphology. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1024–1029. Denver, Colorado: Association for Computational Linguistics.

Boychev, Georgi. 2013. *Case inflection in Koalib: Discovering the rules*. University of Lorraine MA thesis.

Breiman, Leo, Jerome Friedman, Charles J. Stone & Richard Olshen. 1984. *Classification and regression trees*. New York: Taylor & Francis.

Corbett, Greville G. 1991. *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2013. Systems of gender assignment. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner & Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, 1–27. Brussels: Association for Computational Linguistics.

Dimmendaal, Gerrit J. 2015. Accretion zones and the absence of language union. In Gerrit J. Dimmendaal (ed.), *The leopard's spots*, 25–63. Leiden: Brill.

Dowle, Matt & Arun Srinivasan. 2019. data.table: Extension of data.frame. *R package version* 1.12.2. Available at: https://CRAN.R-project.org/package=data.table.

Eddelbuettel, Dirk. 2017. random: True random numbers using random.org. *R package version* 0.2.6. Available at: https://CRAN.R-project.org/package=random.

Gower, John C. 1971. A General Coefficient of Similarity and Some of its Properties. *Biometrics* 27(4). 857–871.

Hammarström, Harald. 2013. Noun class parallels in Kordofanian and Niger-Congo: Evidence of genealogical inheritance? In Thilo Schadeberg & Roger Blench (eds.), *Nuba mountain language studies*, 549–569. Cologne: Rüdiger Köppe.

Hammarström, Harald, Robert Forkel & Martin Haspelmath. 2019. *Glottolog 4.1*. Jena: Max Planck Institute for the Science of Human History.

Kann, Katharina & Hinrich Schütze. 2016. MED: The LMU System for the SIGMORPHON 2016 Shared Task on Morphological Reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 62–70. Berlin, Germany: Association for Computational Linguistics.

Karshola Omar, Hussein, Hassan Komi & Susan Estifanus. 2000. *Riŋeroŋ Kandsagi keḍi Kawaliib [Koalib stories]*. Khartoum: Kwaliib Language Committee.

Kassambara, Alboukadel & Fabian Mundt. 2020. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. *R package version* 1.0.7. Available at: https://CRAN.R-project.org/package=factoextra.

Kaufman, Leonard & Peter Rousseuw. 1990. *Finding groups in data*. New York: Wiley.
Kodi, Ismail. 2000. *Tijaɽina [Traditional Celebration]*. Khartoum: Kwaliib Language Committee.
Krijthe, Jesse. 2018. Rtsne: T-distributed stochastic neighbor embedding using a Barnes-Hut implementation. *R package version* 0.15. Available at: https://github.com/jkrijthe/Rtsne.
Kuhn, Matt & Davis Vaughan. 2019. parsnip: A common API to modeling and analysis functions. *R package version* 0.0.3.1. Available at: https://CRAN.R-project.org/package=parsnip.
Kuhn, Max, Fanny Chow & Hadley Wickham. 2019. rsample: General resampling infrastructure. *R package version* 0.0.5. Available at: https://CRAN.R-project.org/package=rsample.
Kuhn, Max & Hadley Wickham. 2019. recipes: Preprocessing tools to create design matrices. *R package version* 0.1.6. Available at: https://CRAN.R-project.org/package=recipes.
Liaw, Andy & Matthew Wiener. 2002. Classification and regression by randomForest. *R News* 2(3). 18–22.
Maechler, Martin, Peter Rousseuw, Anja Struyf, Mia Hubert & Kurt Hornik. 2019. cluster: Cluster analysis basics and extensions. *R package version* 2.1.0.
Makarov, Peter & Simon Clematide. 2018. UZH at CoNLL–SIGMORPHON 2018 Shared Task on Universal Morphological Reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, 69–75. Brussels: Association for Computational Linguistics.
Milborrow, Stephen. 2019. rpart.plot: Plot rpart models: An enhanced version of plot.rpart. *R package version* 3.0.8. Available at: https://CRAN.R-project.org/package=rpart.plot.
Paluszynska, Aleksandra & Przemyslaw Biecek. 2017. randomForestExplainer: Explaining and visualizing random forests in terms of variable importance. *R package version* 0.9. Available at: https://CRAN.R-project.org/package=randomForestExplainer.
Perry, Patrick. 2017. corpus: Text corpus analysis. *R package version* 0.10.0. Available at: https://CRAN.R-project.org/package=corpus.
Quint, Nicolas. 2006. *Phonologie de la langue koalibe, Dialecte réré (Soudan)*. Paris: L'Harmattan.
Quint, Nicolas. 2010a. Benefactive and malefactive verb extensions in the Koalib very system. In Fernando Zúñiga & Seppo Kittilä (eds.), *Typological Studies in Language*, Vol. 92, 295–316. Amsterdam: John Benjamins Publishing Company.
Quint, Nicolas. 2010b. Case in Koalib (a Kordofanian language) and related Heibanian languages. In *The 40th Colloquium on African Languages and Linguistics*. Leiden: Leiden University.
Quint, Nicolas. 2013. Integration of borrowed nouns in Koalib, a noun class language. In Thilo Schadeberg & Roger Blench (eds.), *Nuba mountain language studies*, 115–134. Cologne: Rüdiger Köppe.
Quint, Nicolas. 2018. An assessment of the Arabic lexical contribution to contemporary spoken Koalib. In Stefano Manfredi & Mauro Tosco (eds.), *Arabic in contact*, 189–205. Amsterdam: John Benjamins.
Quint, Nicolas. 2020. Kordofanian. In Rainer Vossen (ed.), *The Oxford handbook of African languages*, 239–268. Oxford: Oxford University Press.
Quint, Nicolas. 2022. Classes nominales dans deux langues Niger-Congo: le baïnouck djifanghorois (atlantique) et le koalib (kordofanien) [Nominal classes in two Niger-Congo languages: baïnouck and Koalib]. *Faits de Langue* 53. 1–29.
Quint, Nicolas & Siddig Ali Karmal Kokko. 2009. *The phonology of Koalib: a Kordofanian language of the Nuba Mountains (Sudan)* (Grammatical analyses of African languages; Grammatische Analysen afrikanischer Sprachen v. 36 = Bd. 36). Cologne: Rüdiger Köppe. OCLC: ocn517262760.

Quint, Nicolas & Siddig Ali Karmal Kokko. 2022. *Koalib-French dictionary forthcoming*. Paris: L'Harmattan.

R-Core-Team. 2021. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Schadeberg, Thilo. 1981. *A survey of Kordofanian Vol 1: The Heiban group*. Hamburg: Helmut Buske.

Sorokin, Alexey. 2016. Using longest common subsequence and character models to predict word forms. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 54–61. Berlin, Germany: Association for Computational Linguistics.

Suliman, Istifanus. 2000. *Riŋerɔŋw [Stories]*. Khartoum: Kwaliib Language Committee.

Tagliamonte, Sali A. & Harald Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24. 135–178.

Therneau, Terry & Beth Atkinson. 2019. rpart: Recursive partitioning and regression trees. *R package version* 4.1-15. Available at: https://CRAN.R-project.org/package=rpart.

Ting, Kai Ming. 2010. Precision and Recall. In Claude Sammut & Geoffrey I. Webb (eds.), *Encyclopedia of Machine Learning*, 781. Boston, MA: Springer US.

Wickham, Hadley. 2017. tidyverse: Easily install and load the Tidyverse. *R package version* 1.2.1. Available at: https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley. 2019. stringr: Simple, consistent wrappers for common string operations. *R package version* 1.4.0. Available at: https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, Jim Hester & Romain Francois. 2018. readr: Read rectangular text data. *R package version* 1.3.1. Available at: https://CRAN.R-project.org/package=readr.