Quantitative methods in African Linguistics - Predicting plurals in Hausa

Matías Guzmán Naranjo[1], Laura Becker[1]

[1]Universität Leipzig, Germany

Introduction

Despite the availability of electronic resources, African languages have received relatively little attention in the corpus linguistics and computational linguistics community. In this study, we address Hausa plural classes from a computational perspective.

Hausa plurals

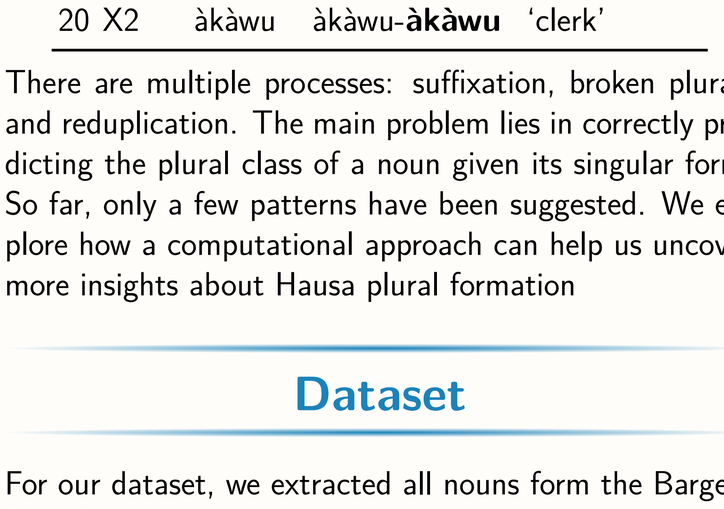
Hausa has a very complex plural system, with anything between 10 and 70 plural classes, depending on how one counts them.
Newman (2000) proposes the following macro-classes, with multiple sub-classes for each one of them.

	Class	Singular	Plural	Gloss
1	a-a	sirdì	siràd**a**	'saddle'
2	a-e	gulbi	gulàb**e**	'stream'
3	a-u	kurmì	kuràm**u**	'grove'
4	-aCe	wuri	wur**àre**	'place'
5	-ai	malàm	malàm**ai**	'teacher'
6	-anni	watà	wàt**ànni**	'moon'
7	-awa	talàkà	talak**awa**	'commoner'
8	-aye	zomo	zom**àye**	'hare'
9	-Ca	tabò	tab**ba**	'scar'
10	-Cai	tudù	tùd**dai**	'high ground'
11	-ce2	ciwò	cìwà**ce-cìwàce**	'illness'
12	-Cuna	cikì	cik**kunà**	'belly'
13	-e2	camfì	càmf**e-càmfe**	'superstition'
14	-i	tàurarò	tàuràr**i**	'star'
15	-oCi	tagà	tag**ogi**	'window'
16	-u	kujèra	kùjèr**u**	'chair'
17	u-a	cokàli	cok**uà**	'spoon'
18	-uka	layò	lay**ukà**	'lane'
19	-una	rìga	rig**unà**	'grown'
20	X2	àkàwu	àkàwu-**àkàwu**	'clerk'

There are multiple processes: suffixation, broken plurals and reduplication. The main problem lies in correctly predicting the plural class of a noun given its singular form. So far, only a few patterns have been suggested. We explore how a computational approach can help us uncover more insights about Hausa plural formation

Dataset

For our dataset, we extracted all nouns form the Bargery (1951) dictionary. The dictionary contains around 3000 nouns, of which only some 1450 have a plural. In the 150 instances where the dictionary provided multiple plural forms for a noun, we only considered the first one for practical reasons.
We assigned all nouns to one of Newman's macro-classes. The distribution of classes can be seen in the following plot. Because some classes have very low frequency, we do not expect to be able to predict these very well.



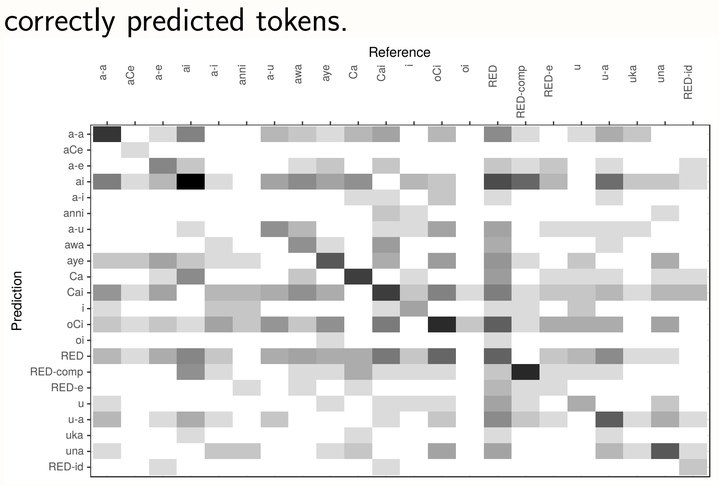
For the predictors, we extracted the final consonant (C.1), the final two vowels (V.1 and V.2), the CV structure of the last four segments (CVCV.4), the length of the singular (length), and the last tone of the singular (T.1).

Statistical method

We use a simple neural network using one hidden layer with five hidden nodes. The network learns the relation between different predictors and the desired outcome. This approach does not let us find out exactly what the relation between the predictors and the outcomes is, but it gives us a powerful model. The formula for the model is: plural.class ~ V.1*T.1 + C.1 + V.2 + CVCV.4 + length The * means an interactions of factors.
We perform ten-fold cross-validation: training the network on 90% of the data and then predicting the remaining 10%. This is repeated 10 times until we have predicted the whole dataset.

Results

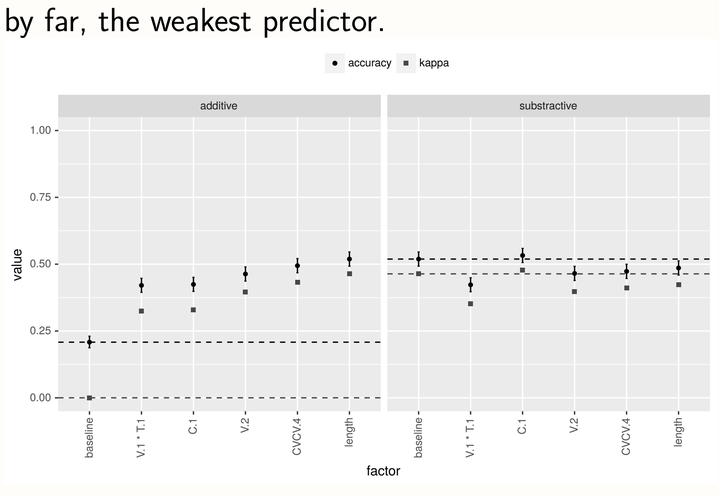
The figure below shows the predicted classes with the attested ones. The correctly predicted classes are in the diagonal; the darker the cell, the higher the number of correctly predicted tokens.



The corresponding statistics are as follows:

Overall Statistics
Accuracy : 0.5425
95% CI : (0.5161, 0.5686)
No Information Rate : 0.2082
Kappa : 0.488

Overall, accuracy is considerably above random chance (similarly with the kappa score).
We can look at the predictive power of each predictor as well. On the left, we have the cumulative accuracy and kappa scores. Those are calculated by starting the model without any predictors, and adding one predictor at a time while calculating the accuracy and kappa scores. This shows us how the model improves with the individual addition of each predictor.
On the right, we see the subtractive accuracy and kappa scores. In this case, the model starts with all predictors, and we compare what happens when only one of the predictors is removed. This reveals how much of a total impact that predictor has (a greater accuracy decrease means the predictor is more important). The most clear point here is that the vowels are much more important for the model than the consonant is. The last consonant is, by far, the weakest predictor.

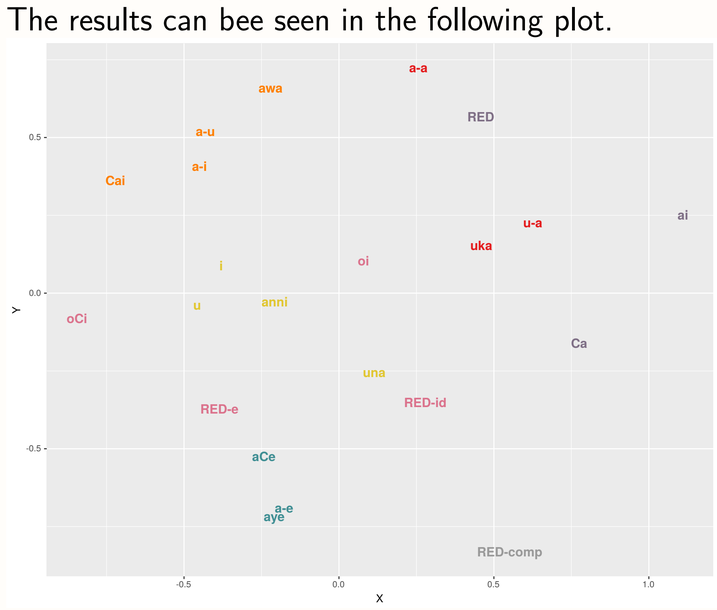


Next, we want to see how 'similar' classes are based on the similarities of their members. To calculate similarities between classes we predict, for each singular form, the probability it belongs to any given plural class (with the same method and cross-validation as before), this method gives us a probability matrix for all classes for all items (on the left).

	a-a	aCe	a-e	ai	...
w_1	0.000	0.000	0.000	0.000	...
w_2	0.059	0.001	0.001	0.001	...
w_3	0.000	0.000	0.012	0.002	...
w_4	0.013	0.016	0.120	0.000	...
w_5	0.000	0.001	0.000	0.012	...
w_6	0.006	0.032	0.018	0.001	...
...					

	a-a	aCe	a-e
aCe	1.509		
a-e	1.575	1.381	
ai	1.663	1.670	1.677

From this later matrix, we calculate the correlation distance (on the right) between the classes ($1 - \mathrm{cor}$). Next, we compress the matrix into two dimensions with multidimensional scaling. This is indicated by the position in the plane. The color coding corresponds to the main clusters (we find these with clustering analysis on the whole similarity matrix).
The results can bee seen in the following plot.



We clearly see that classes that have a similar process of plural creation are closer together on the space, and they also belong together according to the clustering analysis (the color codes). This is strong evidence for the macro classes proposed by Newman.

Conclusions

- Although complex at first sight, Hausa plurals are highly predictable from formal properties of nouns.
- We find strong validation for Newman's macro classes.
- Using computational methods can provide new insights of grammatical systems.
- African languages remain largely unexplored from a corpus and computational perspective, this is a first step towards improving the situation.

References

Bargery, George Percy and Diedrich Westermann (1951). A Hausa-English Dictionary and English-Hausa Vocabulary. Oxford University Press. **Migeod, Frederick William Hugh** (1914). A Grammar of the Hausa Language. K. Paul, Trench, Trübner & co., ltd. **Newman, Paul** (2000). The Hausa Language: An Encyclopedic Reference Grammar. Yale University Press. **Salim, Bello Ahmad** (1981). Linguistic Borrowing as External Evidence in Phonology: The Assimilation of English Loanwords in Hausa. University of York **Schön, James Frederick** (1862). Grammar of the Hausa Language. London: Church missionary house.