An Identity Preference in Ngbaka Vowels[1]

Juliet Stanton
New York University

1 Overview

Models of long-distance phonology differ in whether or not they provide an explicit proviso for identity. In Optimality Theoretic (Prince & Smolensky 2004) terms, this difference is reflected in whether or not a model proposes constraints that reference identity directly.

We can see this difference by considering three proposals that focus on long-distance agreement. On one end of the spectrum, Gallagher & Coon (2009) argue that maintaining identity is a goal that the speaker can have; they propose a constraint IDENTITY, which requires consonants linked by a correspondence relationship to be identical. On the other end, Rose & Walker (2004) and other work in Agreement by Correspondence does not see identity as a goal that the speaker can have, and there are no constraints that reference identity directly. In the middle is Zuraw's (2002) Aggressive Reduplication, where identity is seen as a goal that the speaker can have, but there are no constraints that reference identity directly. In the latter two frameworks, identity is achieved by featural agreement on all possible dimensions; in Optimality Theoretic terms, this amounts to satisfaction of IDENT[±high], IDENT[±sonorant], and so on.

One question raised by this diversity of frameworks is if identity preferences should be analyzed using one monolithic constraint (as in Gallagher & Coon 2009), or if they should be analyzed using a set of constraints, each requiring identity for an individual feature (as in Zuraw 2002, Rose & Walker 2004). This paper presents a case study of Ngbaka vowels that addresses this question through statistical modeling of vowel co-occurrence data. The conclusion emerging from this case study is that both a monolithic identity constraint, as well as constraints that require identity for individual features, are necessary.

2 Background and research question

Ngbaka[2] (Ubangian, Central African Republic and neighboring areas; Maes 1959, Thomas 1963, Henrix et al. 2007, Sélézilo 2008, *a.o.*) has a twelve-vowel system (1). The language has low [ɑ ɑ̃], mid [e ɛ ɛ̃ o ɔ ɔ̃], and high [i u ĩ ũ], which I differentiate with the combination of [±high] and [±low]. The mid vowels contrast for [±ATR]; the nasal vowels missing from the inventory are the nasal equivalents of [+ATR] [e] and [o].

(1) (a) Oral vowels

i		u
e		o
ɛ		ɔ
	ɑ	

(b) Nasal vowels

ĩ		ũ
ɛ̃		ɔ̃
	ɑ̃	

Descriptions of Ngbaka (references above) often claim that multiple types of vowel harmony are active in the language. To give an example, Thomas (1963) acknowledges the presence of an [±ATR] harmony

[1] My thanks to G. Gallagher, M. Gilbert, K. Mooney, NYU's PEP Lab, and the audience at AMP 2021 for feedback.

[2] I use the term 'Ngbaka' here as a cover term for at least two dialects. The Ngbaka discussed by Henrix et al. (2007) and whose lexicon is recorded by Henrix et al. (2015) is distinct from Ngbaka Ma'bo, discussed by Thomas (1963), Sagey (1986), Rose & Walker (2004), and others. I have found no evidence that the vowel patterns of the two dialects differ.

Proceedings of *AMP 2021*

process, evidenced by the fact that [e o] cannot precede or follow [ɛ ɛ̃ ɔ ɔ̃]. She also acknowledges the presence of a backness harmony process amongst non-low vowels, where [i ĩ e ɛ ɛ̃] cannot precede or follow [u ũ o ɔ ɔ̃]. Finally, she acknowledges a height harmony process among back vowels, where [o ɔ] cannot precede or follow [u]. These requirements on vowel co-occurrence often coincide to ensure that, in a CV$_1$CV$_2$ word, V$_1$ and V$_2$ are identical for all features (as noted by Sélézilo 2008).

These facts raise questions about the proper analysis of the preference for identical vowels in Ngbaka. Namely, can the apparent preference for identity in Ngbaka vowels be explained through the interaction of multiple types of vowel harmony, or must an identity preference be recognized independently? Put differently: do we need a monolithic identity constraint to model the patterns, or are IDENT constraints that hold over individual feature values sufficient?

3 Data

To address this question, I created a database of vowel pairs from Henrix et al.'s (2015) Ngbaka dictionary. The dictionary contains a total of 5,571 words of various shapes, but I limited this investigation to disyllabic or longer words that contain only CV syllables (n=3,928). For each word, I extracted each adjacent pair of vowels; examples are in (2-3).[3] Note that the tones are not represented in the vowel pairs; I focus only on segmental similarity and identity here.

(2) z**ì**b**ɔ̀**$_1$l**ɔ̀**$_2$ 'what is right, properly'
Vowel pairs: [i ɔ$_1$], [ɔ$_1$ ɔ$_2$]

(3) t**à**l**ɛ́**k**ù**sì 'type of shrub'
Vowel pairs: [a ɛ], [ɛ u], [u i]

The result was 6,716 vowel pairs. In Table 1, notice that totally identical pairs (n=4,461, in black) are more common than non-identical pairs (n=2,255, in white).

		V$_2$											
		ɑ	**ɑ̃**	**e**	**ɛ**	**ɛ̃**	**i**	**ĩ**	**o**	**ɔ**	**ɔ̃**	**u**	**ũ**
V$_1$	**ɑ**	1322	33	124	76	10	167	6	135	81	14	148	8
	ɑ̃	7	38	1	0	0	9	9	4	3	1	3	1
	e	26	0	239	5	1	5	0	13	7	1	10	0
	ɛ	82	3	9	718	6	17	0	21	24	2	33	0
	ɛ̃	4	4	0	1	35	1	1	2	1	1	3	4
	i	121	6	20	13	2	490	1	59	20	4	10	1
	ĩ	2	2	2	0	0	1	12	0	3	4	0	0
	o	66	2	55	7	1	20	0	532	11	1	18	0
	ɔ	125	10	6	47	3	44	5	16	505	15	25	0
	ɔ̃	2	2	0	4	0	2	5	0	4	18	1	0
	u	184	5	23	29	4	62	3	13	13	1	542	4
	ũ	4	2	0	0	5	1	4	1	0	0	5	10

Table 1: Frequency of vowel pairs from the database

An interesting trend in the table above is that if a vowel co-occurs with a non-identical vowel, one of the vowels in the pair is usually [ɑ]. This point will become relevant later. In addition, there are trends in the data that I don't focus on here (for example: nasal vowels are less common than oral vowels).

4 Analysis

For an analysis, I fit a loglinear model to the count data, using the *bayesglm* function from R's arm package (Gelman & Hill 2007). The baseline model, discussed first, asks the following question: what is the predicted frequency of each vowel pair, given the independent frequency of each vowel, in each position?

[3] Glosses have been translated from the original French by me.

The dependent variable for this model is the number of times a particular vowel-vowel pair is attested. The independent variables in the baseline model include one predictor per vowel per position (ɑ$_1$, ɑ$_2$, ɑ̃$_1$, ɑ̃$_2$, e$_1$, e$_2$, etc.). Table 2 illustrates the structure of the model inputs using a subset of the vowel pairs and predictors.

Vowel$_1$	Vowel$_2$	Combination	Count	ɑ$_1$	ɑ$_2$	e$_1$	e$_2$	i$_1$	i$_2$	o$_1$	o$_2$	u$_1$	u$_2$
ɑ	ɑ	ɑ+ɑ	1322	1	1	0	0	0	0	0	0	0	0
o	i	o+i	20	0	0	0	0	0	1	1	0	0	0
e	e	e+e	239	0	0	1	1	0	0	0	0	0	0
u	ɑ	u+a	184	0	1	0	0	0	0	0	0	1	0
i	o	i+o	59	0	0	0	0	1	0	0	0	0	1

Table 2: Structure of model inputs

The structure of the model was Count$\sim$ɑ$_1$+ɑ$_2$+e$_1$+e$_2$+. . . , and the model was fit using the quasipoisson link function. The predicted counts for each vowel pair, assuming no constraints on co-occurrence, show that identical pairs are overattested relative to expectation (these were obtained with R's *fitted.values* function). This is visible in Figure 1, where pairs that are overattested compared to expectation sit above the identity line and pairs that are underattested sit below it. In Figure 1, a [−ATR] vowel is notated with an <h> and nasality is notated with an <n>; thus <un> corresponds to IPA [ũ], <oh> corresponds to IPA [ɔ], and <ehn> corresponds to IPA [ɛ̃]. (For an interactive plot that allows for examination of the distribution of non-identical vowel pairs, see the author's website.)

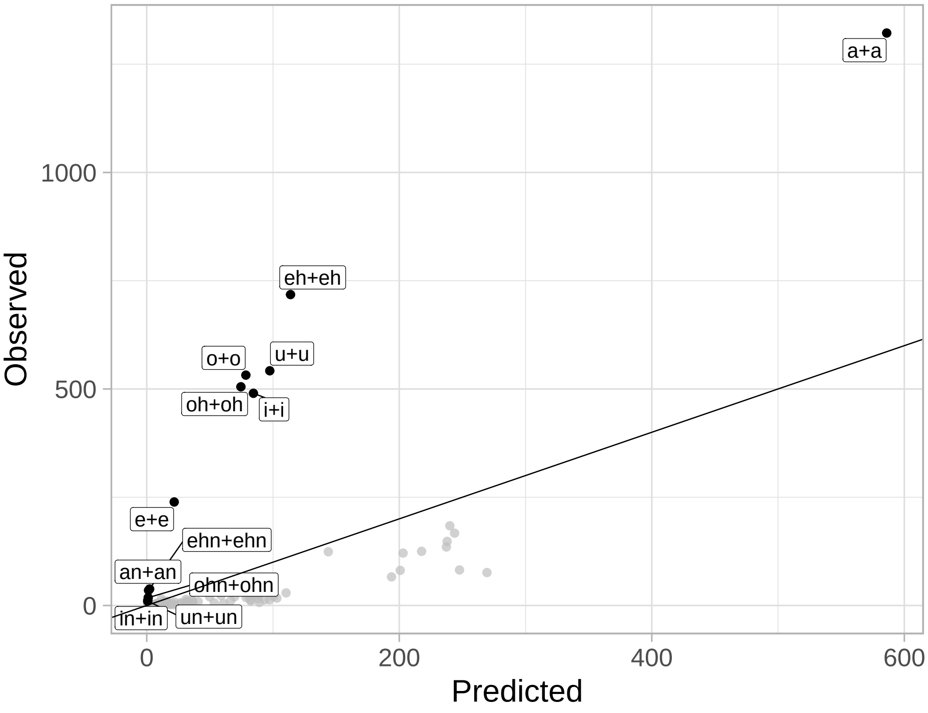

Figure 1: Visualization of model fit

To this baseline model, I added predictors reflecting different types of vowel harmony and one predictor to look for effects of total identity. These predictors and the structures they assign violations to are in Table 3. For convenience of reading, the predictors are formulated as Optimality Theoretic constraints.

These predictors were chosen in large part by consulting published grammatical descriptions of Ngbaka that mention co-occurrence restrictions on vowels (Thomas 1963, Henrix et al. 2007, Sélézilo 2008) and translating the vowel patterns described there into restrictions. As noted above, Thomas (1963) documents [±ATR] harmony, [±back] harmony among non-low vowels, and [±high] harmony among back vowels.

The picture presented by Henrix et al. (2007) differs, in that they document [±high] harmony amongst both [+back] and [−back] vowels (as well as [±ATR] harmony amongst the mid vowels). Sélézilo (2008) does not note any specific forms of vowel harmony but notes a trend towards total identity. Harmony for [±nasal] is not mentioned in any description of Ngbaka that I am aware of; however, trends in Table 1 caused me to consider it for the purposes of statistical modeling.

Predictor (all binary) *Type of harmony*	Assigns a 1 to…
*[αATR][-αATR] *ATR harmony*	(e|o)(ɛ|ɔ) and (ɛ|ɔ)(e|o)
*[αback][-αback] *Backness harmony*	(ie|ɛ)(u|o|ɔ) and (u|o|ɔ)(i|e|ɛ)
*[αhigh, -back, -low][-αhigh, -back, -low] *Height harmony among [−back] vowels*	(i)(e|ɛ) and (e|ɛ)(i)
*[αhigh, +back, -low][-αhigh, +back, -low] *Height harmony among [+back] vowels*	(u)(o|ɔ) and (o|ɔ)(u)
*[αnasal][-αnasal] *Nasal harmony*	vowels mismatching for nasality
*[+syllabic]$_i$[+syllabic]$_j$ *Total identity*	non-identical vowel pairs

Table 3: Predictors added to the model

The relevant question here is: which combination of predictors is responsible for shaping the count data? To answer this question, I fit the maximal model to the data (i.e. the model including all predictors in Table 3) and compared the goodness of fit of nested models using likelihood ratio tests. The models differed only in the presence or absence of the factors in Table 3; all predictors present in the baseline model (see Table 2 for a sample) were present in all of the more complex models.[4]

5 Results

The results suggest that it is necessary to include a constraint that enforces total identity (e.g. Gallagher & Coon's 2009 Identity), in addition to constraints that require identity for individual feature values (e.g. Ident[±ATR]). Thus the inclusion of a constraint that references total identity is justified.

The best-fit model to the data includes the predictors for ATR harmony, Backness harmony among front vowels, Nasal harmony, and Identity. The fact that each of these is significant means that each plays its own independent role in shaping the data. Full model results are presented in Table 4. For the factors referring to harmony: the positive coefficient for Identity means that non-identical pairs are underattested relative to expectation; this is in line with the observation from Figure 1 that identical pairs are overattested relative to expectation. All of the other harmony-based predictors (ATR harmony, Backness harmony, Height harmony among front vowels, and Nasal harmony) have positive coefficients, meaning that mismatches along these dimensions are less frequent than would be expected given the individual frequencies of vowels in first and second position. The individual predictors for each vowel per position follow a similar logic: a positive coefficient means the vowel is underattested, and a negative coefficient means that it is overattested (though very few of these predictors are significant). Of particular relevance to this study is the observation that the total identity predictor has by far the largest coefficient; it plays the largest role in shaping the data.

Further modeling (not represented in Table 3) suggests that another possible type of harmony, between low [ɑ ɑ̃] and the [−low] vowels, is inactive. Put differently, combinations of low vowels and non-low vowels have no explicit penalty associated with them; this is consistent with the observation above that if two non-identical vowels co-occur, one of them is usually [ɑ].

[4] In an ideal world, the model would also incoporate one predictor per vowel combination (a+i, a+ã, e+u, e+ɛ, etc.). This would allow us to gain some insight as to whether or not speakers have generalized beyond restrictions on individual vowel sequences to restrictions on the distribution on feature values. Unfortunately, fitting a model with this many predictors does not appear to be possible using the *bayesglm* function in R.

Predictor	**Coefficient**	***t* value**	**Significant?**
Identity	1.04	29.23	Yes ($p < .001$)
ATR harmony	0.40	5.51	Yes ($p < .001$)
Backness harmony	0.26	4.38	Yes ($p < .001$)
Height harmony among front vowels	0.53	5.07	Yes ($p < .001$)
Nasal harmony	0.63	6.64	Yes ($p < .001$)
ɑ$_1$	-0.68	-2.15	Yes ($p < .05$)
ɑ$_2$	-0.42	-1.39	No
ɑ̃$_1$	0.46	1.36	No
ɑ̃$_2$	0.24	1.36	No
e$_1$	0.06	0.20	No
e$_2$	-0.21	-0.69	No
ɛ$_1$	-0.46	-1.46	No
ɛ$_2$	-0.27	-0.87	No
ɛ̃$_1$	0.41	1.18	No
ɛ̃$_2$	0.35	1.05	No
i$_1$	-0.33	-1.03	No
i$_2$	-0.30	-0.98	No
ĩ$_1$	0.78	2.05	Yes ($p < .05$)
ĩ$_2$	0.38	1.10	No
o$_1$	-0.30	-0.93	No
o$_2$	-0.30	-0.97	No
ɔ$_1$	-0.51	-1.59	No
ɔ$_2$	-0.09	-0.29	No
ɔ̃$_1$	0.66	1.80	Trending ($p = .07$)
ɔ̃$_2$	0.28	0.84	No
u$_1$	-0.48	-1.51	No
u$_2$	-0.20	-0.64	No
ũ$_1$	0.58	1.56	No
ũ$_2$	0.68	1.87	Trending ($p = .06$)

Table 4: Full results of the best-fit model

6 Conclusion

In modeling the count data from Section 2, the fact that the identity predictor is significant tells us that the identity preference in Ngbaka vowels cannot be explained entirely by appealing to interacting processes of vowel harmony. The total identity predictor, in addition to these interacting processes of vowel harmony, plays a role in shaping the data. This finding lends support to theories of long-distance interactions that make explicit reference to identity, e.g. Gallagher & Coon (2009).

References

Gallagher, Gillian & Jessica Coon (2009). Distinguishing total and partial identity: Evidence from Chol. *Natural Language and Linguistic Theory* 27, 545–582.

Gelman, A. & J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Henrix, Marcel, Karel van den Eynde & Michael Meeuwis (2007). *Description grammaticale de la langue Ngbaka. Phonologie, tonologie et morphosyntaxe*. LINCOM Studies in African Linguistics, Munich.

Henrix, Marcel, Michael Meeuwis & Peter Vanhoutte (2015). *Dictionnaire ngbaka-français*. LINCOM, Muenchen.

Maes, Védaste (1959). *Dictionnaire Ngbaka-Francais-Neerlandais*. Commissie voor Afrikaanse Taalkunde / la Commission de Linguistique Africaine, Tervuren.

Prince, Alan & Paul Smolensky (2004). *Optimality Theory: Constraint interaction in generative grammar*. Blackwell, Oxford.

Rose, Sharon & Rachel Walker (2004). A Typology of Consonant Agreement as Correspondence. *Language* 80, 475–531.

Sagey, Elizabeth Caroline (1986). *The representation of features and relations in non-linear phonology*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Sélézilo, Apollinaire (2008). *Le ngbaka manza, langue adamawa-oubanguienne de la RCa: Phonologie, morphophonologie, morphologie et syntaxe*. Omniscriptum, Riga.

Thomas, Jacqueline (1963). *Le parler ngbaka de Bokanga*. Mouton, Paris.

Zuraw, Kie (2002). Aggressive reduplication. *Phonology* 19, 395–439.