CORPUS

Corpus

3 | 2004
Usage des corpus en phonologie

A propos du comportement des verbes trilitères en somali

Xavier Barillot

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/218
DOI : 10.4000/corpus.218
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 1 décembre 2004
ISSN : 1638-9808

Référence électronique
Xavier Barillot, « A propos du comportement des verbes trilitères en somali », *Corpus* [En ligne], 3 | 2004, mis en ligne le 02 décembre 2005, consulté le 08 septembre 2020. URL : http://journals.openedition.org/corpus/218 ; DOI : https://doi.org/10.4000/corpus.218

A propos du comportement des verbes trilitères en somali

Xavier BARILLOT
« Bases, Corpus et Langage »,UMR 6039, UNSA

Résumé : Le but de cet article est double. Il s'agit d'une part de donner une représentation des verbes trilitères (*i.e.* comportant trois consonnes, autrement dit les verbes de forme CvCvC) du somali qui rend compte de l'ensemble de leurs propriétés. D'autre part, au travers de cet exemple d'analyse, je m'attache à discuter du rapport qu'entretient le phonologue avec les corpus, en particulier, leur constitution, leur exploitation, l'appréhension et la résolution des contre-exemples ; on verra aussi que l'analyse proposée (comportement des verbes [CvCvC] du somali) ainsi que ses dividendes, capitaux pour la compréhension de la phonologie de cette langue, auraient été impossibles à obtenir ou du moins non crédibles sans l'utilisation de l'outil informatique. Dans le dernier paragraphe seront discutés les problèmes que posent la constitution de *corpus électroniques* ainsi que leur *exploitation automatique*.

1. Examen de surface des verbes [CvCvC]

On considère les verbes somalis[1] dont la forme de citation (2s impératif) est CvCvC : une liste exhaustive de 329 verbes non dérivés[2] est extraite automatiquement du corpus

1. La langue somalie, parlée dans la corne de l'Afrique, fait partie de la famille couchitique qui appartient au phylum afro-asiatique, au même titre que la famille sémitique.

2. La morphologie somalie est essentiellement suffixale : on a en particulier des morphèmes *-ow*, *-am* et *-an* (respectivement à valeurs inchoative, médio-passive et stative) qui, suffixés à des bases verbales CvC donnent des verbes CvCvC comme *daram* '*s'ajouter*' < *dar* '*ajouter*' ou *duqow* '*vieillir*' < *dúq* '*vieux*'. Pour que la sélection automatique ne contienne pas les formes dérivées, il est nécessaire que cette information soit contenue dans le corpus : il faut que chaque item du corpus possède un attribut *dérivation*. Ceci constitue un premier exemple montrant que le

complet (étudiée depuis la fin du XIX^{ème} siècle, ce qui est relativement précoce pour une langue africaine, la langue somalie dispose actuellement de bonnes descriptions grammaticales et de plusieurs dictionnaires assez importants dont la réunion constitue ce que j'appelle le *corpus complet* : cet aspect sera développé dans les sections 7.1 et 7.2). Ensuite, un examen sommaire des propriétés de cette classe de verbes permet de déceler que cette classe est en fait la réunion de deux classes bien distinctes, qui s'opposent morphologiquement bien que leurs formes de citation soient formellement identiques (CvCvC) : comme on le voit en (1a), la conjugaison de l'une exhibe une alternance entre la deuxième voyelle du radical et rien (*gud**u**b vs gudbaa*), alors que cette voyelle est toujours présente dans la conjugaison de l'autre (*cf.* (1b) : *mat**a**g*, *mat**a**gaa*, …).

(1) Comportement des verbes trilitères du somali

	2s impér.	*1s prés.*	*2s prés.*	*nom verbal*	*glose*
a.	gud**u**b	gudbaa	gud**u**btaa	gudbid	*traverser*
b.	mat**a**g	mat**a**gaa	mat**a**gtaa	mat**a**gid	*vomir*

Sur les 329 verbes, 55% sont de type *gudub* et 40% de type *matag*, le reste hésitant entre les 2 comportements. Cette proportion tend à donner une légère préférence au processus d'alternance voyelle/zéro ; d'ailleurs, sans doute pour cette raison, l'ensemble des linguistes dont Hassan (1994 : 43-45) et Saeed (1993 : 53) qui ont travaillé sur ce sujet s'accordent à considérer la morphologie de *gudub* comme régulière et celle de *matag* comme exceptionnelle.

Le but de ce qui suit est de donner une analyse de ce phénomène : similarité de forme (*gudub* ~ *matag*) pour deux comportements différents (*gudbaa vs. mat**a**gaa*). Plus précisément, je montrerai que cette différence est entièrement prédictible.

---

corpus ne doit pas être une simple liste de mots, mais qu'à l'instar des corpus textuels, on a tout intérêt à l'enrichir au maximum : cet aspect des choses sera discuté dans le paragraphe 7.

Etant donné que la présence ou l'absence d'une voyelle modifie la structure phonotactique du mot en créant des groupes consonantiques, il paraît primordial d'en savoir davantage sur les contraintes qui pèsent sur les clusters de consonnes en somali : c'est pourquoi je propose de mener les deux études suivantes sur le corpus complet.

2. Contraintes sur les groupes consonantiques du somali

La première étude consiste à confirmer ce qu'ont signalé l'ensemble des linguistes ayant travaillé sur le somali dont Saeed (1993 : 18), Cardona (1981 : 11) et Hassan (1994 : 19), à savoir que cette langue refuse les clusters biconsonantiques initiaux et finaux et les clusters triconsonantiques. J'ai ainsi recherché de façon systématique et automatique les clusters consonantiques dans les 40000 entrées du corpus complet : il en ressort que les clusters #CC, CCC et CC# sont *strictement* absents en somali, sauf lorsque la première consonne des groupes CC# et CCC est un glide[3] (*j* ou *w*). La classe des emprunts apporte ici une confirmation éclatante de cette contrainte comme en témoignent les exemples en (2) dans lesquels les clusters interdits en somali sont systématiquement brisés ou simplifiés :

(2) Exemples d'emprunts à l'anglais

t**a**réen	*train*	<	ANGLAIS	**tr**eɪn
ambálas	*ambulance*	<	ANGLAIS	æ**mbj**ʊlə**ns**
inf**u**luwéns**o**	*influence*	<	ANGLAIS	ɪ**nfl**ʊə**ns**
isfántʃ**i**	*éponge*	<	ANGLAIS	**sp**ʌ**nd**ʒ
ist**a**rontíjam	*strontium*	<	ANGLAIS	**str**ɒ**ntj**əm

L'interdiction des clusters #CC, CCC et CC# indique que les seules syllabes autorisées en somali sont CV, CVC,

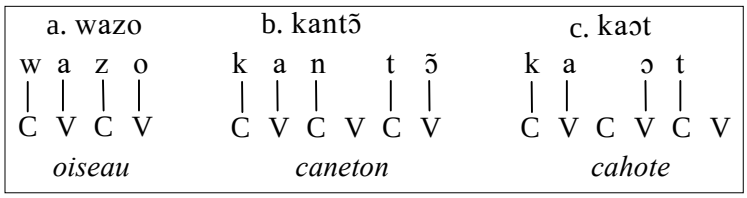
3. Dans ce cas, on peut considérer le glide comme vocalique et non consonantique : ceci est confirmé par la possibilité d'accentuation des glides comme par exemple dans le contraste *éj chien* vs. *ejí chienne* ou *áwr chameau* vs. *awŕ chameaux*.

CVV et CVVC[4], autrement dit que les attaques et les codas branchantes sont exclues[5]. D'après cette contrainte syllabique, la deuxième voyelle des formes *gudub*, *gudubtaa*, *matag* et *matagtaa* est indispensable : son absence créerait des formes impossibles en somali (**gudb*, **gudbtaa*, **matg* et **matgtaa*).

Avant de poursuivre la discussion, je vais présenter rapidement le cadre théorique et le modèle syllabique qui seront utilisés par la suite dans toutes les représentations. Je propose de travailler avec une version simplifiée et très forte de ce cadre, particulièrement adapté à l'analyse des phénomènes du somali qui est une langue sans attaque ni coda branchantes. Pour une discussion concernant des langues moins contraintes au niveau syllabique, je renvoie entre autres à Scheer (1998, 1999).

Dans le modèle syllabique « CVCV » (Lowenstamm 1996), les seuls constituants syllabiques autorisés sont l'attaque et le noyau auxquels ne peut s'associer respectivement qu'une consonne et une voyelle : la ligne squelettale est réduite à une suite de positions consonantiques (C) et vocaliques (V) comme en témoignent les exemples de (3).

(3) Exemples de représentation dans le modèle syllabique « CVCV » (Lowenstamm 1996)



4. Selon la théorie syllabique classique.

5. Cette affirmation réclame des éclaircissements, car les clusters CC intervocaliques du somali peuvent *a priori* être interprétés comme une attaque ou une coda branchante : c'est précisément ce qui serait fait selon le schéma classique dans *habro femmes* où la sonorité du cluster est ascendante. Cependant, selon le procédé de Kahn stipulant que « ne peut être interprété comme attaque branchante que ce qui est attesté à l'initiale de mot », les attaques branchantes n'existent pas en somali. Les emprunts donnés en (2) confirment ceci de façon éclatante : les attaques branchantes sont cassées à l'initiale (*treɪn* > *taréen*) comme en situation interne (*ɪnfluəns* > *infuluwénso*). Par conséquent, en somali, tout cluster biconsonantique intervocalique doit être interprété comme {coda + attaque} : on a *hab-ro* et non **ha-bro*.

Dans la représentation (3b) ci-dessus, il apparaît que certains noyaux sont vides : il s'agit là de la propriété la plus saillante du modèle CVCV ; en fait, contrairement à une théorie syllabique classique, la phonologie du gouvernement suppose qu'un morphème conserve sa structure syllabique, même modifié par un processus phonologique : en particulier, l'alternance V/Ø ne modifie pas cette structure ; le morphème n'est pas resyllabifié, mais c'est juste le contenu d'un noyau qui est modifié : il devient vide (non prononcé). Nous allons voir maintenant dans quelles conditions une position vocalique peut rester vide.

Il s'agit de l'exposé des outils du cadre théorique qui seront cruciales pour l'analyse du phénomène d'alternance V/Ø, c'est-à-dire pour déterminer quand un noyau peut ou ne peut pas être vide. L'impossibilité des formes **gudb*, **gudbtaa*, **matg* et **matgtaa* rejoint un principe très simple de la phonologie du gouvernement (KLV 1990 : 219) selon lequel « toute position vocalique vide doit être proprement gouvernée »[6] (*i.e.* immédiatement suivie par une position vocalique non vide) ; la définition du gouvernement propre et du principe des catégories vides figurent en (4a) et (4b).

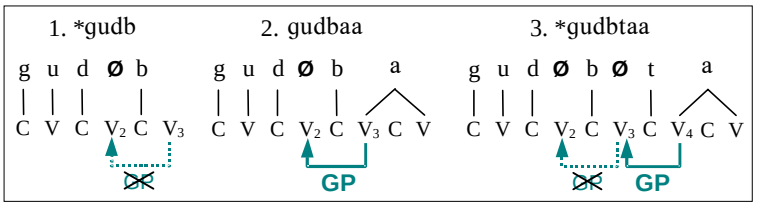
(4) Gouvernement propre et principe des catégories vides

a. Gouvernement propre (GP) :
   Une position vocalique V_1 est *proprement gouvernée* par une position vocalique V_2 ssi (i) V_1 et V_2 sont séparés exactement par une attaque, (ii) V_2 n'est ni vide, (iii) ni proprement gouvernée elle-même.

b. Principe des catégories vides :
   Toute position vocalique vide doit être proprement gouvernée.

c. Représentations de **gudb*, *gudbaa* et **gudbtaa* :

6. Je rappelle que n'est présentée ici qu'une version simplifiée et très forte de la théorie, convenant aux langues sans attaques ni codas branchantes.



La forme *_gudb_ en (4c1) comporte un conflit de gouvernement car V_2 est vide et ne peut être gouvernée par V_2 qui est vide aussi. En (4c3), V_4 gouverne proprement V_3 qui est vide et donc ne peut gouverner V_2 : la forme est mal formée. En revanche, la forme en (4c2) est correcte puisque la position vide V_2 est proprement gouvernée par V_3.

Le gouvernement propre et le principe des catégories vides donnent donc les conditions pour qu'une position vocalique puisse rester vide : « un noyau ne peut être vide qu'à condition que son voisin de droite ne le soit pas ».

Deuxième étude portant sur les contraintes syllabiques du somali, on recherche dans le corpus complet la proportion relative des séquences -VCVCV- et -VCCV- (c'est-à-dire respectivement présente dans _matagaa_ et _gudub_)[7] ; il s'avère que la deuxième est 7 fois plus fréquente que la première : alors que 46% des entrées du corpus complet comportent la séquence -VCCV-, ce chiffre est ramené à seulement 7% pour la séquence -VCVCV-. Ce deuxième résultat nous conduit à considérer que la morphologie régulière des verbes trilitères est effectivement celle de _gudub_ et non celle de _matag_ : la forme _mat**a**gaa_ semble problématique. La prédominance des séquences -VCCV- est un indice pour considérer que, dans le cas des verbes trilitères, la réciproque du principe des catégories vides énoncé en (4b) est vraie : *la deuxième position vocalique des verbes trilitères est vide si elle est proprement gouvernée.*

7. Le corpus complet est limité aux formes de citation (_gudub_ pour les verbes), c'est-à-dire qu'il ne comprend pas les formes fléchies (_gudbaa_, _matagaa_, …) : les verbes CVCVC de ce corpus ne contiennent pas de séquences VCCV et VCVCV. Heureusement, il n'est pas limité aux verbes CVCVC : ce sont les noms et les verbes dérivés qui vont nous renseigner sur ce problème, puisqu'une grande partie d'entre eux est à finale vocalique (ex : _bánqi abondance_ ou _balaqow se ramollir_).

Dans cette optique, les 40% de verbes de type *matag* sont des contre-exemples. Bien sûr, ce résultat de nature statistique ne constitue qu'un indicateur[8] qui ouvre une piste de recherche et que je confirmerai plus loin par l'analyse linguistique.

Afin de comprendre en quoi les verbes de type *matag* sont irréguliers, il est nécessaire de découvrir la représentation des verbes qu'on suppose réguliers, ceux de type *gudub*.

3. Représentation des verbes trilitères de type *gudub*

Etant donné l'alternance *gudub*/*gudbaa*, deux solutions sont logiquement possibles : la forme sous-jacente de ces verbes est soit /CvCvC/ avec les deux noyaux identifiés, soit /CvCøC/ avec un seul noyau rempli. Dans le premier cas, on aura un processus de *syncope vocalique* qui produira [*gudbaa*] à partir de /*gudub*/+*aa* (*cf.* (5a)) et dans le deuxième, un mécanisme d'*épenthèse* qui permettra d'obtenir [*gudub*] à partir de /*gudØb*/ (*cf.* (5b)).

(5) Deux représentations possibles pour *gudub*

a. /gudub/ + aa => gudubaa $\overset{\text{syncope}}{\Rightarrow}$ gudbaa

b. /gudb/ + Ø => *gudb $\overset{\text{épenthèse}}{\Rightarrow}$ gudub

Pour choisir l'une ou l'autre de ces représentations, je propose d'analyser la distribution des deux voyelles des verbes CvCvC du somali.

Toutefois, avant cela, je voudrais faire une remarque qui montre que les choix lexicographiques, loin d'être innocents, sont parfois lourds de conséquence : les deux représentations de (5) ont été proposées dans des langues couchitiques ; la première (*syncope*) a fait l'unanimité en somali comme entre autres dans Bell (1953 : 8), Panza (1974 : 106), Saeed (1993) et Tosco (1997 : 46) et la deuxième (*épenthèse*) a été choisie pour l'oromo, une langue génétiquement très proche

8. Ceci ouvre la question très intéressante et qui sera évoquée au paragraphe 7 de déterminer la réelle signification des résultats statistiques obtenus sur un corpus : de telles statistiques ne peuvent et ne doivent pas répondre à toutes les questions que se posent les phonologues.

du somali qui connaît exactement les mêmes phénomènes d'alternances comme par exemple dans Stroomer (1987 : 55) et Heine (1981 : 21). Pourquoi le même phénomène a-t-il été analysé de deux façons opposées ? Il y a fort à parier que l'explication est simple : la forme de citation des verbes somalis est l'impératif 2s qui comporte toujours les deux voyelles (*gu**u**b* *traverse!*) tandis qu'en oromo, c'est la première personne du présent, ne comportant pas la seconde voyelle (*go**o**nfa* *j'orne*) qui a été choisie comme forme de citation dans les dictionnaires et les grammaires, et ceci bien sûr de manière arbitraire. Ainsi, la plupart des linguistes qui ont travaillé sur ces langues ont considéré à tort les entrées des dictionnaires comme étant les formes sous-jacentes et ont basé sur cet *a priori* toute leur analyse : en somali, une voyelle *disparaît* entre la forme de citation *gud**u**b* et la première personne du présent *gudbaa* (6a) ce qui semble justifier pour certains de postuler une règle de syncope, tandis qu'en oromo, une voyelle *apparaît* entre la forme de citation *gonfa* et la deuxième personne du présent *gon**o**fta* (6b), ce qui donne lieu à une règle d'épenthèse vocalique.

(6) Influence des choix lexicographiques sur l'analyse

		Forme de citation		*Forme fléchie*
a.	SOMALI	gud**u**b	>	1s gudb-aa => SYNCOPE !
		traverse!		*je traverse*
b.	OROMO	gonf-a	>	2s gon**o**f-ta => ÉPENTHESE !
		j'orne		*tu ornes*

Plus qu'un choix lexicographique certes nécessaire mais complètement arbitraire[9] ou en tout cas guidé par des contraintes de nature non phonologiques, c'est l'analyse de la distribution des deux voyelles des trilitères qui va permettre de trancher entre les deux représentations. D'après le tableau (7) dans lequel figure en colonne le comportement morphologique du verbe et en ligne le caractère identique ou non de ses voyelles, les deux voyelles de la grande majorité des 329 verbes

9. Bien sûr, le choix de la forme des entrées de base du corpus complet doit aussi faire l'objet d'un choix réfléchi.

sont identiques : seulement 20% d'entre eux présentent deux voyelles différentes.

(7) Comportement des verbes Cv_1Cv_2C en fonction de l'identité entre v_1 et v_2

Verbes Cv_1Cv_2C	*Voyelles identiques*	**Variante libre**[10]	*Voyelles différentes*	**Total**
v_2 alterne	kidif/kidfaa 159	sabar~sabir/sabraa 11	umal/umlaa 8	*178*
v_2 alterne ou pas	osol/os(o)laa 6	naqal~naqil/naq(i)laa 2	xasif/xas(i)faa 1	*9*
v_2 n'alterne pas	matag/matagaa 91	ʕawar(aa)~ʕawir(aa) 10	dafir/dafiraa 41	*142*
Total	*256*	*23*	*50*	*329*

L'examen de ce tableau révèle une autre propriété très intéressante de ces verbes lorsqu'on croise leur comportement morphologique (alternance ou non) et la nature de leurs voyelles (identiques ou non) : il apparaît que la plupart des verbes à voyelles différentes n'alternent pas et que corollairement les voyelles de ceux qui alternent sont identiques.

En résumé, le tableau (8) ci-dessous fait apparaître les deux propriétés fondamentales des verbes trilitères du somali :

(8) Propriétés des verbes CvCvC en somali

a. La grande majorité des verbes possède deux voyelles identiques.
b. Si la deuxième voyelle d'un verbe alterne avec Ø, il ne possèdera quasiment jamais de voyelles différentes (*i.e. alternance* et *identité des voyelles* semblent être des *propriétés corrélées*).

10. Cette colonne comprend les verbes acceptant deux formes en variante libre, la première avec deux voyelles identiques (*sabar*) et la seconde avec deux voyelles différentes (*sabir*).

Avant de proposer une représentation qui rende compte de ces deux propriétés fondamentales des verbes trilitères, considérons les contre-exemples à la première propriété, relativement nombreux puisque 20% des verbes ont leurs voyelles différentes[11]. Je propose de réduire à néant ce nombre qui, d'après le tableau (7), est tout de même de 73.

Les deux-tiers des contre-exemples apparents partagent une propriété frappante : leur mélodie est *a-i*. Ceux-là seront considérés dans le paragraphe 5. Restent 20 items dont 11 admettent des variantes libres ou dialectales avec deux voyelles identiques comme *silaʕ* ~ *siliʕ* *souffrir*. Par ailleurs, ces 20 verbes ont des finales récurrentes, comme en témoignent les exemples en (9) ci-dessous :

(9) Verbes à voyelles différentes (sauf mélodie *a-i*)

a. qu**a**ʕ *tousser* => abaissement par *ʕ* ou *ħ*.
b. sal**o**w *rugir* => arrondissement par *w*.
c. mur**a**m *se disputer* => abaissement par *m*.

En effet, sur ces 20 verbes, 8 sont à pharyngale finale (en *aʕ#* ou *aħ#*) comme *qufaʕ* en (9a), 4 sont en *ow#* comme *salow* en (9b) et 4 sont en *am#* comme *muram* en (9c) ; ainsi, dans chaque cas, la non-identité des voyelles semble s'expliquer par un effet de la consonne finale.

Avant d'établir le caractère *naturel* et *commun* de ces effets, je me dois de vérifier que cet ensemble de 20 verbes est bien exceptionnel par rapport aux autres verbes trilitères : il faut pour cela comparer la proportion de verbes comportant une pharyngale, un *w* ou un *m* final entre l'ensemble de ces 20 verbes et celui des 309 autres verbes trilitères du somali[12] ; si la

11. Cette proportion doit toutefois être comparée avec ce que le hasard donnerait : si on suppose que chaque voyelle est équiprobable, on aura 75% de chance d'obtenir deux voyelles différentes ; si on donne à chacune d'elles un poids plus « réaliste » (basé sur un dépouillement exhaustif du corpus complet), à savoir 0.5 pour *a*, 0.2 pour *i* et *u* et 0.05 pour *e* et *o*, on obtient alors une proportion de 66.5%, ce qui reste très supérieur aux 20% qu'on a effectivement. C'est donc autre chose que le hasard qui préside au choix des voyelles des verbes trilitères du somali.

12. Je rappelle que le corpus comprend au total 329 verbes trilitères.

proportion s'avère anormalement plus importante dans le cas des 20 verbes à voyelles différentes, ce sera un bon indice pour valider ma décision : leur consonne finale n'est pas choisie au hasard, ce qui explique la non-identité de leurs deux voyelles ; en revanche, si les proportions sont semblables, cela remettra en question ma démonstration. Heureusement, le rapport de proportion est de 4 pour *m* et les pharyngales et de 60 pour *w* : il n'y a que 20% de verbes sans pharyngales, ni *w*, ni *m* en finale dans les 20 verbes tandis qu'ils représentent 85% des 309 verbes. L'effet des pharyngales (abaissement des voyelles moyennes et hautes en *a*) et celui du glide labio-vélaire (arrondissement de *a* en *o*) est relativement commun : il se retrouve de façon massive en somali comme en témoigne les exemples ci-dessous.

(10) Variantes libres impliquant une pharyngale ou le glide *w*

a.	ʕínab	~	ʕánab	*raisin*
	ɖúlluħ	~	ɖúllaħ	*furoncle*
	ħuluúl	~	ħaluúl	*passage étroit*
	gúħud	~	góħod	*herbe rase SP*
	wíʕil	~	wéʕel, wáʕel, wáʕal	*bâtard*
b.	dawlád	~	dowlád	*état*
	gáw	~	gów	*bord*
	t͡ʃáws	~	t͡ʃóws	*noix de muscade*

En (10a) figurent des exemples dans lesquels une pharyngale abaisse une voyelle adjacente ; le dernier exemple est édifiant : il montre le processus d'abaissement de *i* en *a*. Le pouvoir abaissant des pharyngales a été abondamment décrit et discuté, en particulier par McCarthy (1989, 1991, 1994 : 22-24). Ensuite, les exemples de (10b) illustrent le fait qu'en somali, les voyelles *a* et *o* ne sont pas distinctives devant *w*, lorsque ce dernier est en position de coda. Le corpus complet ne comporte aucune paire minimale impliquant les deux séquences *aw* et *ow*, mais contient une petite centaine de paires en variante libre comme les exemples de (10b). Le processus d'arrondissement des voyelles sous l'action des labio-vélaires est d'ailleurs

relativement banal dans les autres langues (Bourciez 1967, Schrijver 1999).

En revanche, l'effet abaissant de la nasale *m* est moins commun : néanmoins, d'après Scheer (1996 : 174-175), « les nasales peuvent abaisser les voyelles » comme c'est fréquemment le cas entre le moyen haut allemand et l'allemand standard actuel. En tout cas, même s'il n'est pas systématique, cet effet est récurrent en somali :

(11) Variantes libres impliquant la nasale *m*

fílgim	~	fílgam	*bâton pointu SP*
dársim	~	dársam	*douzaine*
ʕúrdum	~	ʕúrdam (ʕárdum)	*jeune plante*
mímbar	~	mámbar	*chaire*
mirjaad	~	merjaad, marjaad	*être agité*
gumuuħ	~	gamuuħ	*rendre émoussé*

Le fait que ces trois effets (abaissement par les pharyngales, arrondissement par *w* et abaissement par *m*) soient naturels et se rencontrent couramment dans d'autres langues et surtout en somali[13] permet de considérer que les deux voyelles des 16 verbes trilitères ayant une pharyngale ou *w* et *m* en finale sont sous-jacemment identiques :

(12) Structure profonde des verbes à voyelles différentes

a. qufaʕ < /qufuʕ/ => abaissement par *ʕ* ou *ħ*.
b. salow < /salaw/ => arrondissement par *w*.
c. muram < /murum/ => abaissement par *m*.

Il ne reste que 4 exceptions véritables à la propriété d'identité des voyelles : ils sont reportés ci-dessous en (13).

13. L'extraction automatique des variantes libres contenant *ħ*, *ʕ*, *w* ou *m* montre en effet que ces effets se retrouvent de façon massive en somali, même s'ils ne sont pas systématiques : ils représentent à eux-seuls la moitié (49%) des 770 variantes libres du corpus complet qui impliquent des voyelles radicales. Ce chiffre important est rassurant pour ma démonstration : une grande partie des alternances vocaliques qu'on trouve dans les variantes libres du somali sont dues à l'effet des gutturales, des glides ou des nasales.

(13) Verbes à voyelles différentes : les 4 exceptions

sumad	*marquer au fer*	subag	*étaler du beurre*
umal	*se fâcher*	nidar	*faire un vœu*

Concernant ces 4 verbes exceptions, il est remarquable d'abord que deux d'entre eux comportent la nasale *m* en position médiane, ensuite que *subag* correspond en rendille et en afar (2 langues couchitiques très proches du somali) avec *subaħ* qui a exactement le même sens et enfin que le dernier (*nidar*) non seulement est un emprunt à l'arabe mais admet de plus la variante libre *nidir*. Il ne s'agit donc pas là d'exceptions très « gênantes ».

En résumé, sur les 276 verbes considérés (329 verbes moins 53 verbes en *a-ı*), on n'en trouve que 4 qui n'ont pas leurs voyelles identiques, ce qui ne représente que 1,5% et on peut donc considérer que :

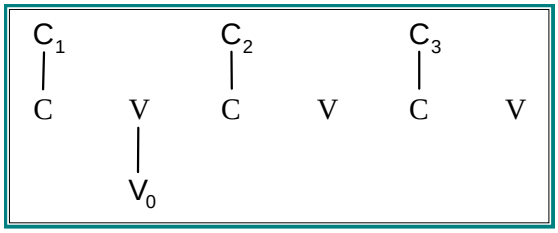
Un verbe trilitère somali a *toujours* deux voyelles identiques

Pour rendre compte de cette propriété fondamentale ainsi que de la propriété (8b) (corrélation entre alternance et identité vocalique), ma proposition est en deux points :

a) La représentation des trilitères est /$Cv_0C\mathbf{Ø}C$/[14] (*cf.* (14)).

b) v_0 se propage sur la deuxième position vocalique lorsque cette dernière n'est pas proprement gouvernée (*cf.* (15a) et (15c)), ce qui a pour conséquence de ne pas promouvoir à la surface les clusters sous-jacents CC# et CCC.

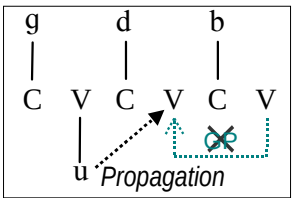
(14) Représentation sous-jacente d'un trilitère $C_1C_2C_3$



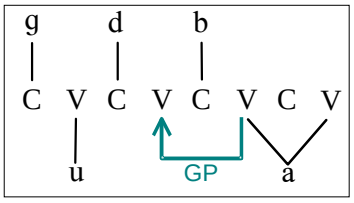
14. La notation Ø indique la présence d'un noyau vide, susceptible d'être rempli dans les conditions schématisées en (15).

(15) Représentation du mécanisme d'alternance V/Ø

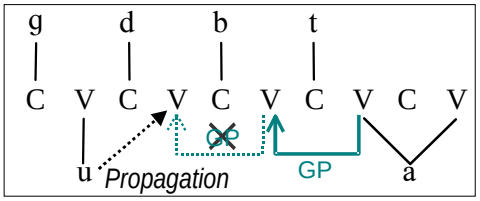
a. 2s impér. : gudub / *gudb



b. 1s présent : gudbaa



c. 2s présent : gudubtaa / *gudbtaa



Le processus de propagation rend compte de manière immédiate de *la systématique identité des deux voyelles* des trilitères.

Par ailleurs, la réalité de ce mécanisme est confirmée par les emprunts trilitères à des mots anglais ou arabes de forme CvCC. Voici en (16) quelques exemples parmi la centaine que j'ai extraits du corpus complet : CvCC n'étant pas viable en somali, la stratégie utilisée consiste systématiquement à copier le contenu du premier noyau dans le deuxième (vide dans la langue source).

(16) Propagation vocalique dans les emprunts CVCC

ANGLAIS	film	>	fílim	*film*
ARABE	baħr	>	báħar	*mer*
	rubʕ	>	rúbuʕ	*quart*
	ʃuʀl	>	ʃúqul	*travail*
	silk	>	sílig	*fil de fer*

Ainsi, pour résoudre des conflits de nature syllabique, la langue somalie utilise de manière courante le processus de propagation, en particulier dans la conjugaison des verbes trilitères (*cf.* (15a) et (15c)). Inversement, lorsque le suffixe

flexionnel est à initiale vocalique comme dans *gudbaa* en (15b), la propagation n'a pas lieu car la forme engendrée ne comporte aucune séquence interdite. La représentation /CvC**Ø**C/ et le mécanisme de propagation (déclenché pour éviter les clusters *CC# et *CCC) rendent compte de manière très naturelle de la propriété (8b), c'est-à-dire du phénomène d'alternance V/Ø.

Cependant, comme je l'ai signalé au début, il existe de nombreux verbes CvCvC en somali qui conservent leur deuxième voyelle en toute circonstance : il s'agit des verbes de type *matag* en (1b) dont la forme 1s présent est *mat**a**gaa* et non **matgaa* comme mon analyse le prédit. L'ensemble de ces 95 verbes, qui représentent 35% des verbes à voyelles identiques, constituent des contre-exemples à ma proposition (14) et (15).

4. Examen des verbes de type *matag*

Que prédit mon analyse pour ces verbes dont la première voyelle se propage *toujours* ? La propagation vocalique, obligatoire dans les formes **gudb* ou **gudbtaa*, n'a lieu qu'en cas de présence de deux noyaux vides successifs (*i.e.* un groupe *CC# ou *CCC) : si tous les noyaux vides sont proprement gouvernés, aucune voyelle n'est propagée. Si ce raisonnement est appliqué à *mat**a**gaa*, cela conduit *inéluctablement* à postuler la présence de deux noyaux vides successifs dans la forme **matgaa*, (autrement dit la présence d'un groupe interdit) : comme dans **gudbtaa* (ou plus précisément *gud**Ø**b**Ø**taa*), ce conflit de gouvernement est résolu par la propagation de la voyelle. Même si en surface, la forme **matgaa* a l'air correcte puisqu'elle comporte un cluster CC tout à fait viable en somali, mon analyse prédit qu'*au niveau sous-jacent*, elle comporte un cluster interdit, en l'occurrence un cluster *CCC :

(17) Structure profonde des verbes non alternants

Malgré les apparences, la forme **matgaa* est /CV**CøCøC**VV/ en structure profonde. Autrement dit, *matag* comprend un cluster central comme dans *hardaf trotter*.

C'est le *mécanisme d'alternance voyelle/zéro*, dont on commence à entrevoir l'importance en somali, qui nous a permis d'énoncer ce résultat crucial.

La phase suivante consiste à donner la représentation exacte de *matag* permettant d'expliquer la présence du cluster *CCC sous-jacent dans **matgaa*. Avant cela, on remarque que ces 95 contre-exemples se partagent en deux classes : les verbes à deuxième et troisième consonnes radicales identiques comme *bara**r** enfler* (au nombre de 43) et les 52 autres comme *matag*.

Commençons par considérer la deuxième classe : la comparaison avec les verbes alternants (type *gudub*) révèle une distribution quasiment complémentaire en ce qui concerne la nature de leur consonne médiane C_2 :

(18) Comportement (+/- alternance) des verbes $Cv_0C_2v_0C_3$ (avec $C_2 \neq C_3$) en fonction de leur consonne médiane

	1a	1b	2	3	
C_2 / Alternance	b, d, ɖ, g, l, m, n, r	ʕ, h, ħ, ʔ	f, s, q	k, t, t͡ʃ, ʃ, x, w, j	Total
+ (*gudub*)	**128**	**17**	**26**	-	171
+/-	3	1	1	1	6
- (*mat**a**g*)	5	4	**9**	**34**	52
Total	136	22	36	35	229

On constate immédiatement dans le tableau (18) que le comportement des verbes est corrélé avec la nature de leur deuxième consonne : il y a *a priori* trois classes de consonnes, celle qui impose l'alternance V/Ø (sous-classes 1a et 1b contenant *b, d, ɖ, g, l, m, n, r, ʕ, h, ħ* et *ʔ*)[15], celle qui l'interdit

15. Les verbes exceptions (verbes non alternants de la classe 1 dans le tableau (18)) sont souvent en variante libre avec des verbes où figure une géminée ou une voyelle longue comme *dabaq ~ da**bb**aq 'couvrir'* ou *tara× ~ t**aa**ra× 'renforcer un ourlet'*. Tout comme les verbes CvCCvC, les verbes CvvCvC ne peuvent pas alterner car le deuxième noyau doit être proprement gouverné pour que le contenu du premier s'y propage : la démonstration et la représentation de ces verbes figurent dans Barillot (2002). Ceci explique pourquoi les verbes comme *dabaq* et *tara×* n'alternent pas. Ensuite, les verbes à gutturale médiane qui n'alternent

(classe 3 : *k, t, tʃ, ʃ, x, w* et *j*) et celle qui autorise les deux comportements (classe 2 composée de *f, s* et *q*). Cette tripartition est à rapprocher (et ce n'est pas un hasard comme nous le verrons plus loin) d'une bipartition basée sur la 'géminabilité' de surface en somali : seules {*b, d, ɖ, g, l, m, n, r*} peuvent géminer phonétiquement. Le tableau (14) permet donc de prédire le comportement de la plupart des verbes CvCvC : un verbe trilitère est non alternant seulement si sa consonne médiane fait partie de {*k, t, tʃ, ʃ, x, w, j, f, s, q*} ; la réciproque n'est vraie que pour les 7 premières consonnes.

Bien sûr, cette distribution a été pressentie par les linguistes qui ont étudié le phénomène (Bell 1953 : 45, Hassan 1994 : 43-45 et Orwin 1995 : 75). Cependant, le fait de ne pas avoir de corpus exhaustif les a empêchés d'une part de construire les trois classes de consonnes (Hassan qui est celui dont l'étude est la plus sérieuse ne cite ni *f, s* et *q*, ni *tʃ* et *ʃ*) et d'autre part de montrer que les verbes CvCvC ont toujours leurs voyelles identiques, ce qui leur interdit de construire une analyse qui rende compte de toutes les propriétés des verbes [CvCvC].

Considérons ensuite la première classe de verbes non alternants, les 43 verbes de type *barar*, c'est-à-dire les trilitères dont les deuxième et troisième consonnes sont identiques : ils sont tous sans exception non alternants, ce qui prouve, comme pour *matag*, qu'ils ont sous-jacemment un cluster /CC/ central. D'autre part, une étude de la distribution de la consonne double donne la restriction étonnante suivante : C_2 appartient presque exclusivement à l'ensemble {*l, r, n, b, d, g*}[16].

La non-alternance des verbes $C_1v_0C_2v_0C_3$ est donc liée aux propriétés segmentales suivantes :

---

pas sont tous du type *sahaj* avec un glide *j* final, ce qui n'est certainement pas un hasard : le somali, pour une raison qui m'échappe, semble rejeter les clusters {*gutturale+j*}. On peut considérer que, canoniquement, les verbes à gutturale médiane alternent.

16. Seuls trois verbes comportent une autre consonne : *f, s* et *j*.

(19) Propriétés segmentales des deux classes de verbes non alternants (*barar* et *matag*)

- type *barar* : $C_2 = C_3$ avec C_2 dans {*l, r, n, b, d, g*}
- type *matag* : C_2 appartient à {k, t, t͡ʃ, ʃ, x, w, j, f, s, q}

Maintenant qu'on sait que les verbes de ces deux classes sont sous-jacemment de la forme /CVC**ø**CVC/ et possèdent donc un noyau vide central, l'étape suivante consiste à trouver et justifier la représentation exacte de *barar* et de *matag* (c'est-à-dire le contenu segmental de la consonne *virtuelle*) ainsi qu'à donner l'analyse qui permet de dériver toutes leurs propriétés, à savoir la non-alternance et les restrictions segmentales énoncées en (19).

Il est possible de montrer que les verbes de type *barar* sont sous-jacemment des réduplicants (/*bar**b**ar*/) et que ceux de type *matag* comportent en structure profonde une géminée centrale (/*ma**tt**ag*/) : la consonne virtuelle est dans le premier cas la consonne initiale du verbe et dans le deuxième, la consonne médiane. En ce qui concerne [*barar*], en plus de l'existence de variantes libres en somali comme *belel ~ belbel brûler*, le comparatisme donne d'excellents indices en faveur de cette hypothèse (par exemple, le mot somali *firir se disperser* est à rapprocher de l'oromo[17] *firfir-sa se disperser*). De même, l'hypothèse concernant *matag* est entre autres appuyée par des considérations comparatistes : dans la plupart des autres langues couchitiques, toutes les consonnes sont phonétiquement géminables (sauf les gutturales) et il est donc très facile de trouver des correspondances comme par exemple entre les noms somali *bakájle lièvre* et afar *ba**kk**eela lièvre*, ou entre les numéraux somali *afár quatre* et rendille *á**ff**ar quatre* ou encore entre les noms somali *ɖoóqo boue* et oromo *ɖo**qq**ee boue*.

Voici en (20) les représentations de *matag* et *barar* dans lesquelles est donnée la nature du cluster CC médian :

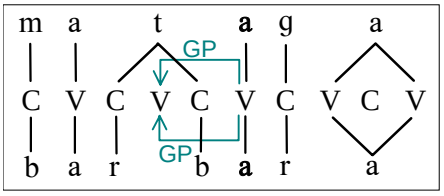
17. L'oromo est une langue couchitique génétiquement proche du somali et parlée dans le sud de l'Ethiopie.

(20) Représentation sous-jacente de *matag* et *barar*

a. La structure sous-jacente de *matag* est /CvC$_i$**ø**C$_i$vC/ : la consonne médiane est sous-jacemment géminée[18].

b. La structure sous-jacente de *barar* est /C$_i$vC$_j$**ø**C$_i$vC$_j$/ : *barar* est en fait un bilitère réduplicqué dans lequel la deuxième consonne C$_i$ n'est pas audible.

c. Schéma des représentations



Bien entendu, les éléments comparatistes donnés ci-dessus ne constituent que des indices et les véritables preuves sont ailleurs, aussi bien pour *matag* que pour *barar* : je n'ai malheureusement pas le temps de présenter la démonstration, trop longue pour être reproduite ici dans son intégralité, ainsi que les nombreux arguments, la plupart élaborés grâce à l'exploitation automatique du corpus informatisé. Je me bornerai à donner rapidement les éléments centraux de la démonstration dont l'intégralité se trouve dans (Barillot 2002) et dans (Barillot & Ségéral *sous presse*).

Commençons par *matag* : je propose de donner l'argument décisif pour considérer qu'un *t* prononcé simple à l'intervocalique est sous-jacemment géminé : prenons le verbe *fur ouvrir*, suffixons-lui le morphème d'autobénéfactif *-(a)t*, puis les morphèmes personnels qui sont *ø* pour 1s en (21a), *n* pour 1p en (21b) et *t* pour 2s en (21c) et enfin le morphème de présent *aa*.

(21) Extrait de conjugaison du verbe autobénéfactif *fur-(a)t*

a.	1s	fur	(a)t	ø	aa	=>	[furtaa]	j'ouvre pour moi
b.	1p	fur	(a)t	n	aa	=>	[furannaa]	nous ouvrons pour nous
c.	2s	fur	(a)t	t	aa	=>	[furataa]	tu ouvres pour toi

18. La consonne est prononcée simple, mais est géminée en structure profonde : ceci s'appelle une géminée virtuelle (Scheer & Ségéral 2001).

La comparaison entre (21a) et (21b) donne les conditions d'apparition du *a* flottant[19] du suffixe autobénéfactif : les conditions sont les mêmes que celles de la deuxième voyelle des verbes trilitères, à savoir que ce *a* n'apparaît que si son absence entraîne la présence de deux noyaux vides successifs. C'est le cas de la forme 1p en (21b), /*furøtøna*/, et le *a* flottant doit donc s'associer pour résoudre le conflit de gouvernement, ce qui donne en surface *furannaa*[20]. En revanche, il n'y a pas de conflit dans la forme 1s, ce qui donne *furtaa* en (21a). Pour 2s en (21c), la forme de surface attendue est *furattaa*, mais c'est *furataa* qu'on obtient : le *t* double est prononcé simple. On pourrait alors penser qu'un des *t* a disparu de la forme sous-jacente, ce qui expliquerait pourquoi un seul est prononcé, mais cette hypothèse est irrecevable : le fait que le *a* flottant soit audible dans la forme 2s nous garantit la présence de deux noyaux vides successifs et donc la présence des deux *t* au niveau sous-jacent (/*furøtøtaa*/). La forme *furataa* comporte une géminée virtuelle. Il est ensuite possible de montrer que tous les *t* (en particulier celui de *matag*) ainsi que les consonnes de l'ensemble {*k, t, t͡ʃ, ʃ, x, w, j, f, s, q*} lorsqu'elles sont prononcées à l'intervocalique sont des géminées virtuelles en somali (Barillot 2002).

La démonstration est plus subtile pour *barar* : je rappelle qu'il faut montrer que *barar* est un réduplicant (/*barbar*/), c'est-à-dire que la consonne virtuelle occupe la position d'attaque dans les modèles syllabiques classiques, ce qui ne manquera pas de surprendre. En revanche, il existe une autre classe de trilitères en somali pour laquelle postuler la même hypothèse ne pose aucun problème : il s'agit des verbes de type *babaʃ* à première et deuxième consonnes identiques. Je suppose pour ces verbes la même représentation sous-jacente que pour *barar*, à savoir /*baʃbaʃ*/[21] : cette fois-ci, la consonne non prononcée est en position de coda ; de plus, les verbes

19. Une voyelle flottante est une voyelle qui alterne avec Ø.

20. Le *t* est assimilé par le *n* : **furatnaa* > *furannaa*.

21. D'ailleurs, sur les 31 mots $C_1VC_1VC_2$ du somali, 30% sont en variante libre avec $C_1VC_2C_1VC_2$ comme par exemple *nanaʃ* ~ *naʃnaʃ bonbon*.

babaʕ contredisent en apparence le principe du contour obligatoire (McCarthy 1986 : 208-211). Cette hypothèse sur la représentation de *babaʕ* est très largement admise par la communauté linguistique et a déjà été proposée pour l'hébreu et le ge'ez par Strelcyn (1948). Or, il s'avère que la consonne finale de ces verbes est en distribution complémentaire[22] avec celle des verbes de type *barar* : cette distribution complémentaire est un argument de poids pour considérer que les deux types de verbes ne forment qu'une seule classe et ont tous les deux la même forme sous-jacente, qui est donnée par *babaʕ*.

Je m'arrête là pour la démonstration qui reste très largement incomplète et dont tous les arguments figurent dans (Barillot 2002). Pour récapituler le résultat concernant les verbes [CvCvC] qui n'alternent pas, je dirais que, dans la représentation en (20c), les deuxièmes *a* de *matagaa* et de *bararaa* sont indispensables car sinon ces formes comporteraient deux noyaux vides successifs. Je rappelle que le fait crucial qui permet d'arriver à cette représentation est le mécanisme d'alternance voyelle/zéro : ce mécanisme constitue la *preuve* de la présence au niveau sous-jacent d'un cluster central dans les verbes *matag* et *barar* ; ceci rend compte du fait qu'ils se comportent de la même façon que *qallaf durcir* ou *daldal lyncher*, comme on le constate ci-dessous en (22) :

(22) Comparaison de *matag* avec *qallaf* et de *barar* avec *daldal*

		2s impératif	*1s présent*	
a.	mat**a**g	mat**a**gaa	/ma**tt**agaa/	*ma**ttg**aa
	qall**a**f	qall**a**faa		*qa**llf**aa
b.	bar**a**r	bar**a**raa	/ba**rb**araa/	*ba**rbr**aa
	dald**a**l	dald**a**laa		*da**ldl**aa

La conséquence de ceci est simple : si les verbes de type *matag* et *barar* se comportent de façon différente de *gudub*, c'est parce qu'il ne s'agit pas de vrais trilitères, mais qu'ils possèdent en fait 4 consonnes dont une n'est jamais

22. C'est en fait un peu plus complexe que cela : la distribution complémentaire repose sur les deux consonnes de ces mots.

audible. La non-alternance V/Ø révèle en effet la présence d'une consonne *quiescente* ou *virtuelle*, la nature de cette consonne étant prédictible d'après la forme de surface :

- si le verbe est [$C_1vC_2vC_2$], la consonne virtuelle est C_1 et la forme sous-jacente du verbe est /$C_1vC_2C_1vC_2$/.
- si le verbe est [CvC_2vC] où C_2 est dans {*k, t, ʃ, t͡ʃ, x, w, j*}, la consonne virtuelle est C_2 et la forme sous-jacente du verbe est /CvC_2C_2vC/.

Avant de continuer, il est crucial de mentionner un dividende très important de ce résultat, à savoir la possibilité pour certaines consonnes du somali de géminer phonologiquement : toutes les grammaires du somali s'accordent à dire que seules sont géminables phonétiquement les occlusives sonores {b, d, ɖ, g} et les sonantes {l, m, n, r}, toutes les autres consonnes ne géminant jamais (Armstrong 1934 : 123, Bell 1953 : 49, 114-115, Panza 1974 : 11, Saeed 1993 : 18 et Orwin 1995 : 6). Ceci est parfaitement exact mais en suivant mon hypothèse (basée sur l'exploitation du corpus des verbes CvCvC) et seulement en la suivant, on est amené à constater qu'à côté de la classe des consonnes phonétiquement géminables, les non géminables se partagent en trois classes : celles qui ne géminent jamais, même phonologiquement, celles qui sont toujours géminées en structure profonde entre deux voyelles et celles qui sont ou ne sont pas sous-jacemment géminées à l'intervocalique. Je récapitule ceci en (23) ci-dessous :

(23) Géminabilité[23] des consonnes en somali : 4 classes

- Consonnes géminables en surface : {b, d, ɖ, g, l, m, n, r} => 1
- Consonnes géminables uniquement en structure profonde :
- {k, t, ʃ, t͡ʃ, x, w, j} : toujours géminées à l'intervocalique => 2
- {f, s, q} : soit géminées, soit simples à l'intervocalique => 3
- Consonnes non géminables : {ʕ, h, ħ, ʔ} => 4

23. Les choses sont un peu plus complexes : en particulier, les occlusives sourdes *k* et *t* ont des contreparties simples à l'intervocalique (réalisées *g* et *d*), tandis que les autres consonnes de la deuxième classe n'apparaissent jamais phonologiquement simples à l'intervocalique. Le cas des glides est particulier lui-aussi. La discussion et les résultats figurent dans (Barillot 2002).

Nous comprenons à présent pourquoi les verbes trilitères dont la consonne médiane est phonétiquement géminable alternent toujours ; si *gudub* n'alternait pas et donnait [*gudubaa*], ce serait que sa consonne médiane serait phonologiquement géminée, ce qui donnerait la forme sous-jacente /*gudØbaa*/, qui ne peut se réaliser autrement que [*guddubaa*] puisque *d* est phonétiquement géminable : le verbe *gudub* doit donc alterner. Seuls les verbes trilitères à consonne médiane non phonétiquement géminable sont autorisés à ne pas alterner.

L'examen du caractère alternant ou non des verbes CvCvC à consonne médiane non phonétiquement géminable nous renseigne sur la géminabilité de cette consonne : ceci amène à poser trois classes de consonnes ; la première concerne les consonnes qui, en position médiane des verbes CvCvC, empêche toujours l'alternance (classe 2) : ces consonnes sont donc toujours phonologiquement géminées à l'intervocalique ([*matag*] < /*matØtag*/) ; la deuxième comprend les consonnes qui autorisent les deux comportements, alternant ou non alternant (classe 3) : elles sont parfois géminée, parfois non en position intervocalique ([*fasaħ*] < /*fasØsaħ*/ *permettre* et [*qasab*] < /*qasØb*/ *contraindre*) ; enfin, la troisième comprend les consonnes qui ne géminent ni phonétiquement, ni phonologiquement.

Je propose à présent, comme annoncé plus haut, d'examiner, à la lumière des résultats concernant *matag*, les verbes en *a-i* comme *dafir répudier* qui sont au nombre de 53 ; je rappelle que ces verbes posent problème à ma représentation des trilitères puisque leurs deux voyelles sont différentes. Le but de ce qui suit est de montrer d'une part qu'ils ne contreviennent pas à mon hypothèse et d'autre part de proposer une analyse qui permet de comprendre leur forme et leur comportement.

5. Les verbes de type *dafir*

La première chose qui intrigue lorsqu'on examine cette liste de près est qu'il s'agit essentiellement d'emprunts arabes : pour seulement 5 d'entre eux, je n'ai pu trouver de

correspondance sûre avec l'arabe. Voilà pourquoi, afin de mieux comprendre les choses, je propose de considérer l'ensemble des mots [CvCvC] du somali empruntés à l'arabe : sur la centaine que contient le corpus complet, la moitié ont leurs deux voyelles identiques et le reste a la mélodie *a-i*. Ce qui est significatif, comme le montrent les tableaux ci-dessous en (24), est la répartition de la consonne médiane entre ces deux classes : alors qu'il n'y a pas de restriction lorsque les deux voyelles du mot somali sont identiques, il ne s'agit que d'une consonne non phonétiquement géminable lorsque la mélodie est *a-i*.

(24) Comportement des emprunts $Cv_0C_2v_0C$ et CaC_2iC en fonction de C_2

Forme du verbe →	$Cv_0C_2v_0C$		CaC_2iC	
Géminabilité de C_2 ↴ *Alternance* →	+Alt	-Alt	+Alt	-Alt
+ géminable phonétiquement $C_2 \in \{b,d,ɖ,g,l,r,m,n\}$	**28**	0	3	0
+/- géminée phonologique $C_2 \in \{f,s,q\}$	**11**	**7**	1	**24**
+ géminée phonologique $C_2 \in \{k,t,x,ʃ,\widehat{tʃ},w,j\}$	0	**15**	0	**25**
Gutturale (non géminable) $C_2 \in \{ʔ,ʕ,h,ħ\}$	**6**	0	0	0

Ainsi, vu l'analyse qui a été proposée pour les verbes de type *matag*, on vérifie que la quasi-totalité des verbes CaCiC sont non alternants : ils ont exactement la même représentation que *matag*, à savoir /CvC_iC_ivC/. Ceci reçoit d'ailleurs une confirmation éclatante lorsqu'on considère les mots arabes dont sont issus ces formes. Voici quelques exemples de correspondances qu'on peut comparer avec des verbes alternants empruntés à l'arabe :

(25) Exemples de verbes CaCiC et CaCaC empruntés à l'arabe

	2s imp.	*1s prés.*	*glose*			
a.	qajir	qaj**i**raa	*changer*	<	ARABE	ʁajjara
	akid	ak**i**daa	*approuver*	<	ARABE	ʔakkada
	fasir	fas**i**raa	*expliquer*	<	ARABE	fassara
b.	dalab	dalbaa	*demander*	<	ARABE	tˁalaba
	safar	safraa	*voyager*	<	ARABE	safara
	qanaʕ	qanʕaa	*être convaincu*	<	ARABE	qaniʕa

Les 3 exemples en (25a) ci-dessus sont très représentatifs de l'ensemble : 75% des verbes arabes dont sont issus les CaCiC du somali sont à la forme II, c'est-à-dire comportent une géminée centrale[24] ; ceci constitue une confirmation de la présence au niveau sous-jacent d'un cluster central dans ces verbes, présence qui a été déduite de leur comportement non alternant. Ainsi, on peut considérer que la structure sous-jacente des verbes [CaC_0iC] est /CaC_0C_0iC/ : pas plus que *matag*, il ne s'agit de verbes trilitères ; ce sont des verbes dont la consonne médiane, non géminable phonétiquement en somali, est géminée phonologiquement. Il paraît logique de se demander ce qui se passe lorsqu'on emprunte un verbe arabe à la forme II dont la consonne médiane est géminable en somali : la prédiction est qu'on devrait avoir la gémination en surface, ce qui est effectivement vérifié comme en témoignent les exemples de (26). Il n'y a d'ailleurs aucune exception : tous les verbes arabes à la forme II dont la consonne médiane est phonétiquement géminable en somali sont prononcés avec la géminée en somali.

24. Les autres verbes somali CaCiC sont principalement issus de verbes arabes à la forme I dont la consonne médiane est toujours phonologiquement géminée à l'intervocalique en somali : elle est donc interprétée comme une géminée comme dans *kaʃif* < ARABE *kaʃafa découvrir*. Pour intégrer un verbe comme *kaʃafa*, la langue somalie a deux possibilités, soit elle respecte la forme de surface et en fait un verbe non alternant (*kaʃif*/*kaʃifaa*), soit elle modifie la forme de surface pour produire un verbe alternant, par exemple en remplaçant le *ʃ* par un *s*, ce qui donnerait *kasaf*/*kasfaa*. C'est la première stratégie qui est choisie.

(26) Emprunts CvC_iC_ivC à la forme II de l'arabe

faddil	*favoriser*	<	ARABE	fadˁdˁala
qallib	*tourner*	<	ARABE	qallaba
baddal	*changer*	<	ARABE	baddala

Les verbes de type *dafir* ne sont donc pas de vrais trilitères comme le sont les verbes de type *gudub* : comme ceux de type *matag*, leur représentation sous-jacente est équivalente à celles de *faddil* ou *baddal*.

Lorsqu'un verbe [CvCvC] n'alterne pas (donc que sa structure sous-jacente est /CvCCvC/), on a deux possibilités pour sa mélodie vocalique, soit les voyelles sont identiques (*matag*), soit elles sont *a-i* (*dafir*). La présence de *a* dans *dafir* s'explique aisément car elle apparaît aussi à cet endroit en arabe. En revanche, la présence *i* est plus problématique ; il y a *a priori* deux possibilités pour l'expliquer : soit cette voyelle provient de l'arabe, soit elle est une création du somali ; nous allons voir que la réponse se situe au milieu.

La forme II est invariablement CaCCaCa en arabe classique : elle ne contient jamais la voyelle *i*. Pour trouver la trace du *i*, il faut considérer le dialecte arabe égyptien[25] : j'ai effectué une extraction *exhaustive* des verbes à la forme II à partir du dictionnaire arabe égyptien - anglais de Badawi & Hinds (1986) ; ceci m'a permis de constituer un corpus informatisé de 1471 formes CvCCvC. Les résultats laissent apparaître deux classes de verbes : 55% ont la forme CaCCaC et le reste CaCCiC. L'exploitation de ce corpus met en évidence que le choix entre ces deux classes est piloté par la nature des deux dernières consonnes radicales (donc celles qui sont adjacentes au *i*) :

25. Les principales influences arabes sur le somali proviennent de l'arabe classique par l'intermédiaire de l'écrit (religion, droit, commerce, …), mais aussi et surtout d'Egypte et du Yémen en ce qui concerne l'oral.

(27) Distribution de CaCCaC et CaCCiC au sein de la forme II en arabe égyptien

a. Le verbe est $X_1\alpha X_2X_2\alpha X_3$ lorsque X_2 ou X_3 font partie de l'ensemble {ʔ, ʕ, h, ħ, x, ʁ, q, $z^ʕ$, $s^ʕ$, $t^ʕ$, $d^ʕ$, r}[26].
b. Le verbe est $X\alpha X_0X_0\iota X$ dans tous les autres cas.

La présence du *a* entre X_2 et X_3 est sans surprise pour les 11 premières consonnes de l'ensemble de (27a) : l'activité abaissante des gutturales et des emphatiques est bien connu. En revanche, cette activité est moins commune pour la liquide *r*.

Revenons au somali : la comparaison avec l'arabe classique et l'arabe égyptien permet de différencier quatre classes parmi les verbes de type *dafir* dont voici des exemples :

(28) Emprunts somalis à la forme II de l'arabe

	SOMALI	AR. ÉGYPTIEN	AR. CLAS.	
a.	sallaħ	sˁallaħ	sˁallaħa	*ajuster*
	fakar	fakkar	fakkara	*penser*
b.	labbis	labbis	labbasa	*habiller*
	akid	ʔakkid	ʔakkada	*approuver*
c.	baddal	baddil	baddala	*changer*
	a͡tʃal	ʔaggil	ʔaʒʒala	*différer*
d.	sarrif	sˁarraf	sˁarrafa	*échanger*
	qajir	ʁajjar	ʁajjara	*changer*

Supposons que les voyelles des verbes somalis sont issus des verbes arabes : en (28a), rien ne permet de savoir *a priori* si la langue source est l'arabe classique ou égyptien ; en revanche, en (28b) et (28c), on a toutes les raisons de penser que les mots sont respectivement issus de l'arabe égyptien et de l'arabe classique. La catégorie (28d), qui est loin d'être négligeable puisqu'elle représente presque le tiers des emprunts

26. Broselow [1976 : 139-140] constate une distribution équivalente pour la voyelle des verbes de la forme I qui est *a, i* ou *u* de façon non prédictible, sauf lorsque C_2 ou C_3 font partie de l'ensemble (21a) où elle est invariablement *a*.

du somali à des verbes arabes à la forme II, pose problème et remet en question l'hypothèse selon laquelle la mélodie vocalique de ces mots est issue de l'arabe : la mélodie vocalique de ces verbes est “innovée”. Les 20 verbes de cette catégorie figurent ci-dessous :

(29) Emprunts somalis à vocalisme « innové »

SOMALI	AR. ÉGYPTIEN	AR. CLAS.	
axir	ʔaxxar	ʔaxxara	*retarder*
bajid	bajjadˁ	bajjadˁa	*blanchir*
dafir	naffar	naffara	*répudier*
fasil	fasˁsˁal	fasˁsˁala	*confectionner*
fasir	fassar	fassara	*expliquer*
kawir	kawwar	kawwara	*enrouler*
xasif	qasˁsˁaf	qasˁafa	*briser*
xasir	xassar	xassara	*gaspiller*
laqin	laqqan	laqqana	*dicter*
naʃir	naʃʃar	naʃʃara	*étendre*
nawir	nawwar	nawwara	*blanchir*
qajir	ʀajjar	ʀajjara	*changer*
sawir	sˁawwar	sˁawwara	*peindre*
ħaqir	ħaqqar	ħaqqara	*mépriser*
ħajir	ħajjar	ħajjara	*embarrasser*
jasir	jassar	jassara	*faciliter*
ʕaddil	ʕatˁtˁal	ʕatˁtˁala	*gêner*
debbir	dabbar	dabbara	*arranger*
qarrir	qarrar	qarrara	*déterminer*
sarrif	sˁarraf	sˁarrafa	*échanger*

Un examen sommaire de cette liste permet de remarquer (1) que les trois-quarts de ces verbes ont la liquide *r* en finale ou en C_2 et (2) que le reste des verbes a soit une emphatique, soit une uvulaire en C_2 ou C_3 dans le mot arabe. Ceci explique pourquoi on a une mélodie *a-a* en arabe égyptien. Or, si ces consonnes ont pour effet d'abaisser le *i* en *a* en arabe égyptien, ce n'est pas le cas en somali : on a vu plus haut en (9a)

et (10a) que seules les pharyngales *ʕ* et *ħ* (et *m* dans une moindre mesure) ont ce pouvoir dans cette langue. D'ailleurs, les verbes somalis /CvC$_0$C$_0$vC/ empruntés à l'arabe et comportant une pharyngale en C$_2$ ou C$_3$ ont *tous* la mélodie *a-a*.

La voyelle *i* n'est donc pas empruntée à l'arabe égyptien ou classique : le mécanisme est plus complexe. Devant ce problème, je propose l'analyse suivante : pour une raison que j'ignore, les verbes arabes à la forme II ne sont introduits en somali qu'avec leur première voyelle, qui est toujours un *a* (*i.e.* `arabe` *fassara* > `somali` /fasøsør/) ; la langue somalie doit ensuite résoudre le problème de la présence des deux noyaux vides successifs. Selon mon hypothèse, la solution consiste à faire une épenthèse de *i* au niveau du deuxième noyau vide. Signalons d'abord que *i* est la voyelle épenthétique du somali ; voici quelques exemples d'épenthèse par *i* dans des emprunts :

(30) Exemples d'épenthèse par *i*

wáqt**i**	*temps*	<	ARABE	waqt
isfánt͡ʃ**i**	*éponge*	<	ANGLAIS	spʌnd͡ʒ

D'après cette hypothèse, en somali, tous les verbes à géminée centrale, qu'elle soit ou non réalisée, devraient avoir la mélodie *a-i*. Sur les 151 verbes /CvC$_2$C$_2$vC/ du somali (empruntés ou non), il y a 57 exceptions. Cependant, comme cela a été mentionné plus haut, parmi ces exceptions figurent les verbes ayant une gutturale en C$_3$ dans lesquels le *i* est systématiquement abaissé en *a*, comme par exemple dans *sallaħ ajuster* (/*saløløħ*/ > **salliħ* > *sallaħ*) ou *majaʕ faire un caprice* (/*majøjøʕ*/ > **majjiʕ* > *majaʕ*) : ces verbes, qui ne posent pas problème puisqu'on connaît l'effet abaissant des pharyngales en somali, sont au nombre de 34. Le reste des exceptions est marginal : il ne représente que 15% des 151 verbes[27].

27. Ce chiffre descend à 9% si on ne considère que les emprunts arabes de la forme II : il s'agit souvent de verbes très courants, probablement empruntés à l'arabe classique au tout début du contact entre les somalis et les arabes (*fakar penser*, *baddal changer*, …).

Mon hypothèse est donc que la voyelle *i* des verbes CaCiC n'est pas empruntée à l'arabe égyptien, mais que le somali a mis en place un *mécanisme d'épenthèse*. Ceci peut paraître surprenant : le somali dispose dans les verbes arabes de deux voyelles que rien n'empêche d'intégrer et pourtant il ne garde que la première et invente pour remplir la deuxième un processus d'épenthèse. Je n'ai pas grand chose à dire concernant cette invention[28]. Cependant, un argument fort en faveur de l'épenthèse est que les verbes /$Cv_1C_iC_iv_2C$/ ont soit leurs deux voyelles identiques, soit la mélodie *a-i* : aucune autre mélodie n'est attestée[29].

La représentation des verbes à géminée centrale du somali est donc /CvC_0C_0**Ø**C/ avec la plupart du temps une épenthèse de *i* (comme dans *ɖabbis lisser* ou *wajig hésiter* qui ne sont pas des emprunts arabes) éventuellement abaissé par une pharyngale (comme dans *ɖabbaʕ s'allonger* ou *sajaħ se couvrir de rosée* qui ne sont pas non plus d'origine arabe) et beaucoup plus rarement une propagation de la voyelle (comme dans *qallaf* ou *matag*).

6. Conclusion

Au terme de la discussion des verbes à mélodie *a-i*, il apparaît qu'il n'y a plus d'exception[30] à la définition des trilitères

28. Elle permet au somali de satisfaire le *principe de récupération* (Barillot 2002) : ce principe permet la récupération de la forme sous-jacente d'un verbe et donc de son comportement (alternance ou pas) à partir de la considération de sa seule forme de surface ; si la mélodie vocalique d'un verbe est *a-i*, il s'agit d'un verbe à géminée centrale et sinon, ce verbe est un vrai trilitère ; par exemple, malgré l'ambiguïté caractérisant *s*, on sait que le verbe *basar résoudre* est trilitère, tandis que *ʃasir* expliquer est quadrilitère.

29. A part les quelques exceptions apparentes, à savoir les verbes dont une des voyelles a été abaissée par une pharyngale (*niggaħ sceller*, *muddaʕ discuter* ou *ʕabbur suffoquer*), il n'y a que trois véritables exceptions dont deux sont des emprunts arabes (*arrun faire accroupir*, *jirrab tester* et *sabbur patienter*).

30. En plus des 4 verbes alternants recensés en (8d), il faut rajouter les 4 verbes en *a-i* de la troisième colonne du tableau (26) qui alternent : il faut cependant signaler que, parmi ces quatre verbes, deux admettent une variante à voyelles identiques (*sabir ~ sabar patienter* et *daris ~ daras étudier*) et un admet une variante qui n'alterne pas (*xasif / xasfaa*

donnée en (14) et (15). Le corpus initial des verbes [Cv_1Cv_2C] se divise en deux classes ; la première est véritablement trilitère et possède les caractéristiques suivantes : représentation sous-jacente /CvCøC/, alternance V/Ø et $V_1 = V_2$; la deuxième est constituée des quadrilitères de type *dafir* (*matag*) et *barar*, qui n'alternent pas, la deuxième voyelle étant le plus souvent un *i* épenthétique dans le cas de ceux de type *dafir*.

Cette étude nous a permis de mettre en évidence des dividendes très intéressants et inattendus pour la phonologie du somali, en particulier une redéfinition complète de l'ensemble des consonnes géminables, la mise en place d'un mécanisme d'épenthèse par *i* pour les quadrilitères. Cela a permis aussi et surtout de découvrir le *rôle-clef* que joue le *mécanisme d'alternance voyelle/zéro* dans toute la langue : il prouve de façon incontestable la présence d'une consonne virtuelle dans les verbes CvCvC de type *matag* et *barar*.

La partie phonologique de cet article est à présent terminée. Cependant, je souhaite consacrer un développement à l'élaboration et l'exploitation des corpus, capitaux dans mon travail comme on a pu le constater au cours de cette étude. L'essentiel de mes recherches sur corpus, des vérifications et arguments destinés à appuyer mes hypothèses reposent sur l'utilisation de l'informatique : parmi les résultats qui précèdent, certains n'auraient pas vu le jour, d'autres seraient restés approximatifs, non seulement sans un bon cadre théorique et sans de bonnes intuitions de départ, mais aussi sans *l'informatisation du corpus complet* et sans *les outils d'interrogation* que j'ai élaborés. L'objet du paragraphe suivant est précisément d'exposer les méthodes de constitution et d'investigation de corpus électroniques.

7. Annexe : Généralités, outils et méthodes

7.1 Introduction : spécificités d'un corpus de phonologue

De façon consensuelle, un *corpus* est un ensemble d'énoncés que le linguiste extrait de l'ensemble *infini* des

~ *xasif/xasiifaa bouleverser*). Le dernier est en fait la seule vraie exception (*daris être voisin*) parmi ces 4 verbes en *a-i* qui alternent.

énoncés possibles pour mener son étude ; un corpus doit constituer un échantillon représentatif de la langue ou des aspects de la langue qu'on veut étudier et doit être fini.

Bien que rien n'interdise a priori de faire de la phonologie sur de tels corpus, qu'on peut appeler des *corpus textuels*, les phonologues utilisent plutôt des *listes d'items non ordonnés*, extraites d'un ensemble *fini*, à savoir le lexique de la langue en question avec éventuellement toutes les formes fléchies, ce que j'appelle *corpus de base* ou *corpus complet*. Par conséquent, le problème de la représentativité d'un corpus extrait ne se pose pas de la même manière en phonologie et en analyse de discours : on n'a jamais accès au corpus complet en analyse de discours alors qu'on peut l'approcher presque autant qu'on veut[31] en phonologie.

En fait, le volume et la nature du corpus de base dépendent du niveau de connaissance de la langue. S'il s'agit d'une première description, le corpus sera relativement restreint par rapport au lexique complet et, pour se donner les moyens de réaliser l'analyse phonémique dans les meilleures conditions, devra être transcrit avec la plus grande précision. Inversement, si la langue a déjà été abondamment décrite et si par exemple on dispose de dictionnaires, le corpus de base s'approchera de l'exhaustivité et les items pourront être retranscrits phonologiquement (il suffira d'avoir présent à l'esprit les règles de réalisation).

Dans la suite de ce paragraphe, je suppose que le phonologue dispose d'un corpus de base quasi exhaustif : le problème de la représentativité ne se pose que si la taille du corpus et le moyen d'extraction interdisent de récupérer toutes les données pertinentes ; dans ce cas, le phonologue devra se résigner à extraire du corpus complet un échantillon *représentatif* des données qui l'intéressent : il lui faudra faire des hypothèses sur la structure de la langue pour choisir, en fonction de l'étude qu'il souhaite mener, le sous-ensemble du corpus complet dont il extraira les données.

31. Le phonologue est limité par la qualité des dictionnaires de base, la compétence des informateurs avec lesquels il travaille et l'existence de mots non répertoriés, néologismes, emprunts, argot, acronymes, etc.

7.2 Constitution du corpus complet

La première question qui se pose est l'accès au corpus de base : on admet ici qu'on dispose de dictionnaires relativement complets. En somali par exemple, le corpus de base que j'ai constitué comprend les entrées des quatre dictionnaires (Keenadiid 1976), (APS 1985), (ZOL 1991) et (Farah 1995) qui possèdent respectivement 14500, 36000, 27500 et 12000 entrées dont une partie a été soit confirmée, soit infirmée par un informateur[32]. Au total, ceci donne un ensemble hétérogène : chaque dictionnaire peut contenir des erreurs (on obtient donc des conflits), est structuré selon des choix lexicographiques différents (forme de citation différente, notation des pluriels nominaux dans des entrées à part ou non, …) ou se base sur une aire dialectale sensiblement différente. La constitution du corpus complet pose donc un premier problème qui est celui de son hétérogénéité : il faut impérativement pour chaque item et chaque information concernant cet item indiquer la source avec si possible un degré de confiance ; nous verrons plus loin des exemples d'entrées du corpus complet.

Le deuxième problème concerne les limites du corpus complet : qu'est-ce qui fait partie du corpus d'une langue pour un phonologue ? En particulier, les emprunts[33], les formes fléchies, les morphèmes dépendants, les formes peu usitées, les mots iconiques[34], … doivent-ils être intégrés au corpus ? La

32. On peut raisonnablement estimer que ces quatre dictionnaires regroupent plus de 95% des mots utilisés ou compris par les locuteurs somalis les plus érudits : en témoigne le nombre dérisoire de mots donnés par mes informateurs qui n'y figurent pas.

33. Bien sûr, il n'est pas toujours facile de décider qu'un item est ou non un emprunt : un mot somali dont la forme et la glose sont semblables à celles d'un mot arabe n'est pas forcément un emprunt, car cette similarité peut aussi être due à l'origine commune afro-asiatique de ces deux langues. En théorie, l'emprunt a pu aussi se produire dans l'autre sens, mais c'est très improbable dans le cas de l'arabe et du somali.

34. On peut se demander comment savoir de façon sûre si tel item est iconique ou non : l'intuition du linguiste, celle de l'informateur et l'information fournie par les dictionnaires ne sont pas forcément cohérentes sur ce point (par exemple, d'après mon informateur, le mot *×ıβ* '*silence*' est iconique).

réponse appartient à chaque phonologue, mais je pense qu'il y a intérêt à intégrer le maximum d'items au corpus, l'essentiel étant de bien spécifier pour chacun d'eux s'il s'agit d'emprunt, de mot iconique, … : pour qu'un corpus soit au maximum « réutilisable », il est indispensable que rien n'en soit exclu, mais que chaque entrée soit clairement identifiée par un code, de façon à ce qu'un utilisateur quelconque puisse facilement extraire les données qui l'intéressent. Pour ma part, je n'ai rejeté que les morphèmes dépendants et les formes fléchies, mais toutes les formes dérivées ont été conservées. Le choix, qui peut surprendre, de ne pas conserver les formes fléchies est justifié par le type de recherche que je mène : la plupart des phénomènes que je cherche à expliquer ne procèdent pas de la phonologie pure mais de la *morphophonologie*. Je m'explique : lorsqu'on cherche à établir des paires minimales ou à déterminer quels clusters de consonnes sont exclus, il est préférable, sinon indispensable dans un souci d'exhaustivité de considérer aussi les formes fléchies ; en revanche, en morphophonologie, on s'intéresse à la forme des morphèmes, à la relation qu'ils entretiennent avec la forme de base : on a juste besoin de la liste des formes de base, avec pour chacune une description de son comportement morphologique. En d'autres termes, pour le phonologue, une forme fléchie ou une forme de base ont très souvent le même statut dans le corpus, tandis que pour le morphophonologue, chaque forme fléchie est reliée à sa forme de base : on a une structure à plat dans le premier cas et une structure hiérarchique dans le deuxième[35].

35. Le choix de ne pas conserver les formes fléchies en somali est aussi justifié par le temps et le volume de données : pour chaque verbe, il y a environ 70 formes conjuguées différentes (plus de 150 si on compte les formes homonymes) et pour chaque nom, il y a au minimum 5 formes fléchies (pluriel et déclinaisons). Etant donné que le corpus somali est constitué de 10000 verbes et de 30000 noms, sa taille, et avec elle le temps de saisie, aurait été 20 fois plus importants ! Comme ceci sera précisé ci-dessous, l'absence des formes fléchies du corpus est compensée par une description complète des propriétés de chaque item qui permet de reconstruire chaque forme fléchie, à condition, en ce qui concerne les verbes par exemple, de disposer d'une ‘conjugaison-type’.

Ceci nous amène à un point important qui est l'*informatisation des corpus de base*. Dans les paragraphes précédents et dans la discussion ci-dessus, on a vu que le corpus n'est pas seulement une liste nue d'items, mais qu'il s'avérait nécessaire d'affecter à chaque item, une série d'informations ou d'*attributs* comme leur *provenance*, un *degré de confiance*, un *type de comportement* (morphologique), mais aussi leur *catégorie*, leur *glose*, leur *statut* dans la langue (mot dérivé, emprunt, dialecte, iconicité, marginalité, …), la *transcription API* si la correspondance avec la graphie n'est pas régulière, …

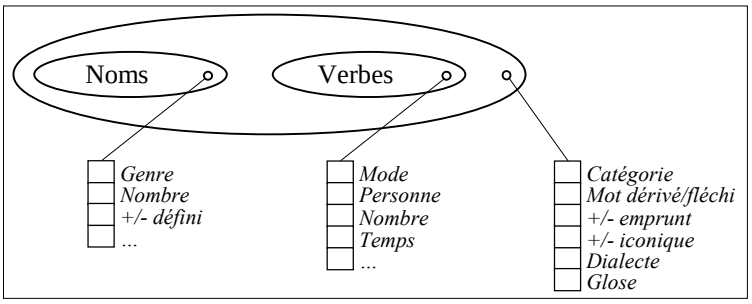
Un grand nombre d'informations doit donc être affecté à chaque item du corpus ; si ce dernier a de plus une taille conséquente, son exploitation devient extrêmement fastidieuse, surtout si par exemple on cherche à croiser des informations ou à comparer le nombre de paires minimales entre couples de segments. La solution naturelle consiste à utiliser l'outil informatique afin de constituer un *corpus électronique*. Ce travail nécessite deux phases :

- élaborer une modélisation des données et le format de saisie qui en découle.
- saisir l'ensemble des données.

Commençons par la deuxième phase, la saisie : il s'agit bien sûr d'un travail extrêmement gourmand en temps ; à titre d'exemple, la saisie du dictionnaire d'oromo de Gragg (1982) comportant 8100 entrées avec catégorie, glose et divers renseignements (pluriel pour les noms, statut d'emprunt, sous-catégorisation pour les verbes, …) a nécessité environ 65 heures de saisie : bien sûr, ce chiffre devra être révisé en fonction du nombre d'attributs, des éventuelles modifications à apporter à la graphie des entrées, etc. La saisie est cependant loin de ne constituer qu'une phase pénible et sans intérêt : elle permet d'abord une grande familiarisation avec les structures de la langue ; à son terme, on est en mesure d'avoir de bonnes intuitions. Ensuite, elle permet de mesurer la confiance qu'on peut avoir dans les dictionnaires pour tel type de données. Bien sûr, un corpus informatisé comporte toujours des fautes de saisie, mais la plupart d'entre elles sont facilement détectables automatiquement : d'abord, on peut créer un outil vérifiant que

les entrées sont toujours classées dans l'ordre alphabétique et ensuite sont conformes à la structure syllabique de la langue (en somali par exemple, on n'a jamais trois consonnes à la suite, ni deux consonnes initiales ou finales, ni encore deux voyelles différentes adjacentes).

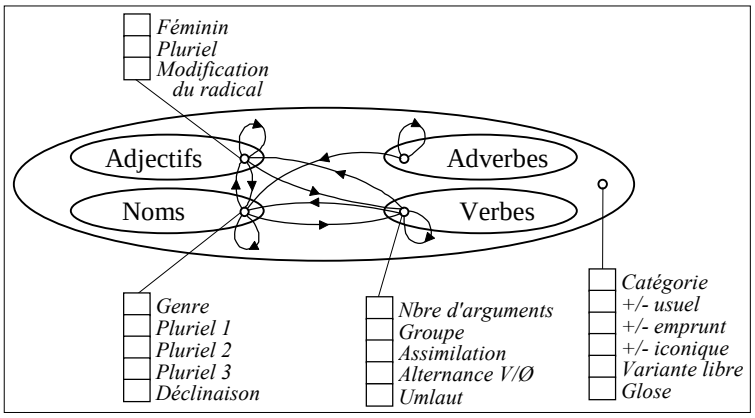
Le choix de la modélisation des données[36] dépend à la fois de la langue en question et du type d'extraction qu'on veut faire. Comme je l'ai signalé plus haut, si on ne fait que de la phonologie pure, la modélisation sera relativement simple : elle pourra se limiter à une classe d'objets (les items) avec une série d'attributs que le phonologue choisira en fonction des recherches qu'il souhaite mener. Voici un exemple de modélisation de ce type pour les noms et les verbes :



En revanche, si la morphologie joue un rôle dans les phénomènes étudiés, il peut être extrêmement intéressant, afin de faciliter ou même de permettre l'extraction automatique d'objets ayant certaines propriétés, de se construire un schéma de données plus complexe, par exemple pour exprimer le rapport de dérivation entre items ; la base de données aurait alors une structure proche de celle d'un dictionnaire étymologique. Ce type de modélisation va par exemple permettre d'extraire facilement les verbes trilitères admettant à

36. Elaborer un *modèle de données* consiste à décrire l'organisation des données : chaque donnée, qu'on peut appeler un objet, a des propriétés ; ces propriétés permettent de définir des *classes*, constituées d'objets qui ont des *attributs* (propriétés) communs ; il est ensuite possible de définir des *liens* entre classes, de regrouper des classes dans des *hyperclasses*, etc. Nous verrons plus bas des exemples simples de modèles ou *schémas* de données.

la fois un nom verbal féminin en *-aal* et un dérivé causatif en *-sii*, ce que la première ne permettait pas sans intervention manuelle fastidieuse. Voici ci-dessous un exemple de modélisation simplifiée pour la langue somalie dans laquelle chaque flèche signifie « dérive de ».



On remarque dans ce schéma que la description des formes fléchies (verbale et nominale) apparaît en tant qu'attribut des objets contrairement au premier schéma dans lequel ces formes étaient des objets : la classe *verbes* ne contient que la forme de citation des verbes, chacun d'entre eux étant ensuite décrit par les valeurs des attributs de la classe, en particulier leur nombre d'arguments, mais aussi s'ils alternent ou pas (*gudub*/*gudbaa*), si leur voyelle subit un umlaut lorsqu'on suffixe le morphème flexionnel *i* de l'infinitif (*jab*/*jebi casser*), etc. A partir de ces attributs et de la conjugaison complète d'un verbe-type, il est possible de construire toutes les formes fléchies de tous les verbes. En revanche, dans le premier schéma, la classe *verbes* contient toutes les formes fléchies, d'où la nécessité d'avoir les attributs *mode*, *personne*, *nombre*, *temps*, etc. qui spécifient la flexion dont il s'agit ; ces attributs n'ont pas lieu d'être dans le deuxième schéma.

Bien sûr, le modèle de données décrivant la morphologie d'une langue est bien plus complexe, mais pour le type de requête qu'un phonologue est amené à faire, le deuxième schéma sera très souvent amplement suffisant. Dans

le cas contraire, nous verrons dans la section suivante qu'il est toujours possible d'élaborer des outils d'interrogation de la base.

Une fois le corpus de base constitué en base de données, plus ou moins complexe selon les besoins du linguiste, nous passons à son exploitation, c'est-à-dire à l'extraction de listes d'items ayant certaines propriétés et qui serviront de base pour établir des hypothèses sur le fonctionnement de la langue.

7.3 Exploitation du corpus

Le raisonnement *infra* suppose l'existence d'un corpus électronique presque exhaustif, suffisamment riche pour qu'on n'ait pas besoin de consulter les dictionnaires. Ce corpus est en même temps 'plat', c'est-à-dire que n'y figurent pas les liens entre classes déterminant les relations de dérivation : ces informations sont contenues dans la base, mais en tant qu'attribut du mot dérivé ; la construction de la base de données avec les liens peut donc se faire de manière automatique, mais n'a pas été réalisée pour le moment. Voici ci-dessous un court extrait du corpus complet du somali :

```
aafo.nf6m^8(aafaad.ka)~désastre, sévère dommage<ARA
_aafee.v2e-tr~faire du tord à, endommager
_[I]aafeyn.nvf:aafee
aagaan.nf30m~récipient haan (petit récipient à lait)
ag.nmcol~alentours, environs, voisinage
arag.v1sc-tr~voir
_aragbeel.v1n-it-CMP~devenir aveugle, perdre la vue
°baad.v1nd-it~passer la nuit dans un endroit inconnu
bakaal.nm^0=bakool~planète Venus\TUN:bakaal\SID:bakkalca
(kala)bixi.v2-tr~séparer (plusieurs personnes)
bitijor.nm1f~lampe torche<N<FRA
canaan.nf^1~réprimande, injures<ARA:anna~dire des injures
[I]dhalaal.v1n-it~circuler (sang)
```

De multiples informations apparaissent dans cet extrait comme le *dictionnaire source* ([I], ^), la *catégorie* (v, n), la *conjugaison* (1, 2), le *genre* (f, m), le *type de pluriel* (1, 3, 8, …), le *timbre vocalique* (°), les emprunts (<ARA, <FRA, …), les mots dérivés (_), les noms verbaux (nv), la glose (~), l'alternance V/Ø des verbes (s, n), les rapprochements avec d'autres langues (\), …

A présent, on va chercher à exploiter ce corpus, c'est-à-dire à en extraire des informations automatiquement : cela réclame l'élaboration d'un logiciel d'interrogation. La plupart des bases de données commercialisées comme DbaseIII ou Access possèdent leur propre langage de requête, mais compte-tenu de la spécificité des requêtes qu'on est amené à faire en phonologie, j'ai choisi d'élaborer un logiciel d'interrogation que j'ai appelé *MBase*. Ce logiciel permet d'extraire un ensemble d'objets en fonction de la valeur de leurs attributs, mais aussi en fonction de leur forme : par exemple, on peut choisir de ne sélectionner que les objets de type CVCCVVC (comme *ʕadʕeéd soleil* en somali) ou commençant par CaC- ou encore les objets contenant deux consonnes identiques séparées par une voyelle longue (-C_iV$V C_i$-), etc. Le programme[37] permet de définir des digraphes et s'adapte à la plupart des corpus puisqu'il n'impose aucun format de saisie : c'est l'utilisateur qui spécifie les séparateurs (entre attributs), ce qui implique qu'il doit bien connaître la structure du corpus.

Il reste cependant des informations, très spécifiques aux phonologues, que le logiciel *MBase* ne peut extraire du corpus. C'est le cas en particulier de la recherche des paires minimales : pour que *Mbase* soit capable d'extraire les paires minimales du corpus, il faudrait enrichir le modèle de données d'un lien de *ressemblance* entre items. J'ai donc élaboré un autre programme, spécifique à l'extraction des paires minimales et que j'ai appelé *Pmin*. Ce logiciel demande deux chaînes de caractères *a* et *b* à l'utilisateur, extrait les mots M_a contenant *a* et les mots M_b contenant *b* et renvoie tous les couples (*Ma*, *Mb*) pour lesquels *Ma* est égal à *Mb* si on y remplace *a* par *b*. *Pmin* permet donc par exemple de trouver les paires minimales impliquant *a* et *e*, mais permet aussi de trouver tous les couples de mots ne différant par exemple que par *ard* et *ald* pour savoir si *r* et *l* sont distinctifs entre *a* et *d* (le corpus somali ne comporte qu'un seul couple : *q**ald**án mauvais* / *q**ard**án solide*) ou encore les couples ne différant que par *man* et *ho* comme

37. Ecrit en Turbo Pascal, il comporte environ 1000 lignes d'instructions.

saman temps / *saho frotte!*[38] ; de plus, on peut spécifier la position dans le mot, le contexte de chaque chaîne de caractères et la catégorie des mots qui forment la paire minimale. Encore une fois, sans le logiciel *Pmin* et sans informatisation du corpus, il est inutile de penser à chercher de façon systématique les neutralisations ou le rendement des oppositions entre phonèmes. D'autres programmes ont été écrits, comme la mise en correspondance lexicale entre langue qui compare plusieurs corpus (calcul de distance entre items basée sur la similarité de forme qui peut être définie par l'utilisateur et basée facultativement sur la similarité de sens) ou comme la recherche de cooccurrences consonantiques dans des formes de base de la langue.

Ce dernier exemple fournit l'occasion d'évoquer un sujet dont il n'a pas encore été question : la ***validation*** d'un résultat obtenu grâce à l'exploitation d'une partie du corpus. Toutes les recherches ne se prêtent pas à cette méthode de validation, contrairement aux recherches de cooccurrences. Prenons l'exemple suivant : en arabe et dans la plupart des langues afro-asiatiques, on observe des restrictions entre les consonnes d'une racine (McCarthy 1979, 1981) ; à titre d'exemple, *b* et *f* ne sont jamais adjacents. En fait, les consonnes de même lieu d'articulation s'excluent, ce qui amène à les répartir dans des classes. J'ai effectué une recherche systématique sur les 1700 mots bilitères du somali (*CVC* et *CVVC*) pour déterminer si cette langue est soumise à des restrictions analogues pour ses 22 consonnes, et pour éventuellement établir les classes d'équivalence de la relation d'exclusion. Un logiciel simple permet d'obtenir le tableau 22x22 des cooccurrences entre consonnes et j'ai ensuite défini manuellement les classes de consonnes qui s'excluent. La méthode de validation consiste ensuite à confronter cette hypothèse (la répartition en classes) sur une autre partie du corpus, par exemple sur les 1000 trilitères : dans le cas où le nombre de mots contenant deux consonnes d'une même classe

38. A proprement parler, il ne s'agit pas, pour ces deux derniers exemples, de recherche de paires minimales, puisque plus d'un son est impliqué dans chacun des deux éléments.

(définie à partir des bilitères) reste en deçà d'une valeur raisonnable, l'hypothèse sera validée et nous pourrons conclure que le somali est soumis à des restrictions de cooccurrence, comme l'arabe.

7.4 Conclusion

L'utilisation de l'informatique, qui demande un gros investissement au départ, a un intérêt énorme : comme nous l'avons vu dans l'exemple de la morphologie des verbes [CVCVC], j'interroge constamment le corpus électronique, au départ bien sûr pour constituer la liste d'items qu'on va étudier et à partir de laquelle on va bâtir des hypothèses sur la langue, mais aussi très souvent au cours de la démonstration, soit pour trouver des arguments, soit parce qu'il s'avère qu'une partie des items du départ partage des propriétés avec d'autres, etc. Sans l'utilisation de l'outil informatique, rien de tout ceci n'aurait été possible et on aurait été réduit bien souvent à des conjectures.

L'informatisation des corpus et de leur exploitation autorise un type de recherches qu'il est impensable de faire manuellement : ce que l'ordinateur fait en quelques minutes (recherche des paires minimales dans un corpus de 40000 entrées) de façon bête demanderait au minimum plusieurs centaines d'heures à un linguiste qui travaille de façon intelligente. Le deuxième avantage de l'outil informatique est la possibilité d'obtenir des résultats *exhaustifs* : les propriétés qu'on déduit de l'interrogation d'un corpus électronique complet ne sont pas discutables ; cela permet de baser ses analyses, ses hypothèses *sur des données propres*.

Il est utile, toutefois, de se prémunir contre certains écueils propres aux corpus informatisés : d'abord, le choix des lexicographes n'a souvent rien de phonologique comme on a pu le constater sous (6) ; ensuite il faut être prudent avec les chiffres : une certaine proportion ou un nombre d'exceptions doit être interprété en fonction du contexte dans lequel il a été obtenu, en particulier du type d'exception, du nombre total d'items du corpus extrait, etc. ; enfin, l'informatique est souvent extrêmement utile (parfois même indispensable), mais n'est pas un but en soi : *l'informatique doit rester un outil*. En aucun cas

ne doit-elle se substituer au cadre théorique et à l'intuition du linguiste, indispensables à l'interprétation des données.

Références bibliographiques

APS = Agostini F., A. Puglielli & C.M. Siyaad & *alii* (1985). *Dizionario somalo-italiano*. Roma : G. Gangemi Editore.

Armstrong L.E. (1934) : « The phonetic structure of Somali », in *Mitteilungen des Seminars für orientalische Sprachen zu Berlin*, 37-3 : 116-161. [Reprint *The phonetic structure of Somali* (1964). East Ridgewood, New Jersey : Gregg Press].

Badawi E.-S. & M. Hinds (1986). *A dictionary of Egyptian Arabic (Arabic-English)*. Beirut : Librairie du Liban.

Barillot X. (2002). *Morphophonologie gabaritique et information consonantique latente en somali et dans les langues est-couchitiques*. Université Paris VII (thèse de doctorat).

Barillot X. & P. Ségéral (sous presse). « On phonological processes in the '3rd conjugation' of Somali », in R.J. Hayward, J. Ouhalla & D. Perret (éds.) *Studies in Afroasiatic grammar*. Amsterdam and Philadelphia : John Benjamins publishing company.

Bell C.R.V. (1953). *The Somali language*. London, New York and Toronto : Longman, Green and Co.

Bourciez E & J. Bourciez (1967). *Phonétique française. Étude historique*. Paris : Klincksieck.

Broselow E.I. (1976). *The phonology of Egyptian Arabic*. PhD Thesis, University of Massachusetts.

Cardona, G.R. (1981). « Profilo fonologico del Somalo », in G.R. Cardona & F. Agostini eds *Studi Somali 1 - Fonologia e Lessico*. Roma : Ministero degli Affari Esteri, Direzione Generale per la Cooperazione allo Sviluppo, 3-26.

Farah A.A. (1995). *Abwaan cusub oo af-Soomaali iyo af-Ingiriisiya (A modern Somali-English dictionary)*. Ottawa : A.A. Farah « Barwaaqo ».

Gragg G.B. (1982). *Oromo Dictionary*. East Lansing : African Studies Center, Michigan State University.

Hassan M.M. (1994). *Aspects de la phonologie et de la morphologie du somali*. Thèse de doctorat, Université de Nice-Sophia Antipolis.

Heine B. (1981). *The Waata dialect of Oromo : Grammatical sketch and vocabulary*. Berlin : Dietrich Reimer Verlag.

Keenadiid Y.C. (1976). *Qaamuuska Af-Soomaaliga*. Firenze : E. Ariani.

KLV = Kaye J., J. Lowenstamm & J.-R. Vergnaud (1990). « Constituent structure and government in phonology », *Phonology Yearbook* 7-2 : 193-231.

Lowenstamm J. (1996). « CV as the only syllable type », in J. Durand & B. Laks (eds) *Current Trends in Phonology : Models and methods*, vol. 2. Salford : ESRI, 419-441.

McCarthy J.J. (1979). *Formal problems in Semitic phonology and morphology*. Thèse de doctorat, MIT [Publiée en 1985, New York : Garland].

McCarthy J.J. (1981). « A prosodic theory of non-concatenative morphology », *Linguistic Inquiry* 12, 3 : 373-418.

McCarthy J.J. (1986). « OCP effects : Gemination and antigemination », *Linguistic Inquiry* 17, 2 : 207-263.

McCarthy J.J. (1989). *Guttural phonology*. Amherst : University of Massachusetts.

McCarthy J.J. (1991). « Semitic gutturals and distinctive features theory », in B. Comrie & M. Eid (eds) *Perspectives on Arabic linguistics III*. Amsterdam & Philadelphie : John Benjamins publishing company, 62-91.

McCarthy J.J. (1994). « The phonetics and phonology of Semitic pharyngeals », in P. Keating (ed.) *Papers in laboratory phonology III*. Amherst : University of Massachusetts, 191-283.

Orwin M. (1995). *Colloquial Somali. A complete language course*. London & New York : Routledge.

Panza B. (1974). *Af Soomaali. Grammatica della lingua somala con piccolo vocabulario in appendice*. Firenze : Le Monnier.

Pillinger S. & L. Galboran (1999). *A Rendille Dictionary Including a Grammatical Outline and an English-Rendille Index*. Cologne : Rüdiger Köppe Verlag.

Parker E.M. & R.J. Hayward (1985). *Afar-English-French Dictionary (with Grammatical Notes in English)*. London : SOAS, University of London.

Saeed J.I. (1993). *Somali Reference Grammar*. Kensington, Maryland : Dunwoody Press.

Scheer T. (1996). *Une théorie de l'interaction directe entre consonnes : contribution au modèle syllabique CVCV. Alternances e-ø dans les préfixes tchèques, structure interne des consonnes et la théorie X-barre en phonologie*. Thèse de doctorat, Université Paris VII.

Scheer T. (1998). « Governing domains are head-final. Structure and Interpretation. Studies in Phonology », in E. Cyran (ed) *Studies in phonology*. Lublin : Folium, 261-285.

Scheer T. (1999). « A theory of consonantal interaction », *Folia linguistica* 32 : 201-237.

Scheer T. & P. Ségéral (2001). « Abstractness in phonology : the case of virtual geminates », in K. Dziubalska-Kolaczyk (ed) *Constraints and Preferences*. Berlin et New York : Mouton de Gruyter, 311-337.

Schrijver P. (1999). « Vowel rounding by primitive Irish labiovelars », *Ériu* 50 : 133-137.

Strelcyn S. (1948). « Les racines trilitères à première et deuxième radicales identiques », in *Comptes-rendus du GLECS*, tome IV, années 1945-48 : 88-89. Paris : Ecole pratique des hautes études, Sorbonne.

Stroomer H. (1987). *A Comparative Study of Three Southern Oromo Dialects in Kenya : Phonology, Morphology and Vocabulary*. Hamburg : Helmut Buske Verlag.

Tosco M. (1997). *Af Tunni : Grammar, Texts, and Glossary of a Southern Somali Dialect*. Köln : Rüdiger Köppe Verlag.

ZOL = Zorc, R.D., M.M. Osman & V. Luling (1991). Somali - English Dictionary. Kensington, Maryland : Dunwoody Press.