NEGATION ET INTONATION EN KABYLE.

Amina METTOUCHI

Cet article d'hommage reprend en partie certains éléments présentés à une journée sur l'intonation en berbère à laquelle Naïma avait participé, en mai 2004 au Centre de Recherche Berbère, à l'INALCO.

Elle y avait présenté des données berbères de l'oasis de Siwa issues d'un travail en cours avec Gérard Philippson, et nous avions discuté des phénomènes d'accentuation de la négation 'la'. Nous avions parlé de faire ensemble une étude comparative de la négation portant sur différentes langues berbères. Le destin en a décidé autrement, et cette collaboration amicale et professionnelle n'aura pas pu voir le jour. Néanmoins, chaque fois que j'aborde les questions de prosodie dans ma recherche, le souvenir de Naïma est présent et m'accompagne, comme je suis sûre qu'il accompagne beaucoup de mes collègues.

Introduction

Le point de départ de ce travail est la conclusion d'une recherche en morphosyntaxe et énonciation, portant sur la valeur de l'élément postverbal de la négation, *ara*, en kabyle. La présente étude intonative vise à analyser un paramètre prosodique, la fréquence fondamentale (F0), qui nous est apparu perceptivement comme essentiel dans l'interprétation des valeurs co-énonciatives prises par les énoncés négatifs, ainsi que dans la construction des repérages interpropositionnels.

La cadre théorique dans lequel nous nous situons pour cet article est celui de la grammaire de l'intonation, mis en place par Morel et Danon-Boileau (1998) pour le français, et en cours d'extension à d'autres langues. Dans ce cadre, les variations de F0 jouent un rôle primordial dans la co-énonciation, c'est-à-dire la façon dont celui qui parle se représente et anticipe la pensée de l'autre. Selon les auteurs : « une montée intonative marque iconiquement un appel à l'attention de l'autre. Elle permet de forcer son attention sur un 'thème' ou sur un focus, enjeu d'un possible désaccord. La plage basse au contraire marque un retour sur soi. » (1998 : 15). A côté des valeurs iconiques, les variations du fondamental ont également des valeurs conventionnalisées. « *La première*

est sans doute la valeur de continuatif de la remontée de F0 en fin de segment. C'est une façon comme une autre de forcer l'attention de l'autre en lui manifestant que l'on n'a pas fini de s'exprimer. Ces remontées permettent secondairement de constituer des sous-ensembles au sein des constituants discursifs majeurs inscrits entre deux pauses silencieuses. Contrairement aux silences, les montées établissent une articulation entre les éléments qu'elles séparent. » (1998 :16). Même si ces valeurs ont été dégagées à partir du français, elles se retrouvent dans d'autres langues (voir par exemple Magro (2003) pour le maltais) et nous ferons l'hypothèse que les plus générales d'entre elles sont valides pour nos données. On retrouve des hypothèses similaires (bien qu'exprimées dans d'autres cadres théoriques) dans Wichmann (2000) à propos de l'anglais, et dans Hirst et Di Cristo (1998) pour différentes langues.

Nous commencerons par synthétiser les conclusions auxquelles nous sommes parvenue suite à l'analyse morphosyntaxique et énonciative des énoncés négatifs, puis nous présenterons les profils intonatifs des différents types, en fonction de deux critères : présence ou absence de l'élément postverbal *ara*, et syntaxe simple ou complexe de l'énoncé.

1. Synthèse de l'analyse morphosyntaxique et énonciative.

Dans nombre de descriptions du kabyle, la négation est présentée comme un morphème discontinu, composé de l'élément négatif proprement dit (*ur* ou l'une de ses variantes), et d'un élément de renforcement négatif postverbal (*ara* ou l'une de ses variantes). On aurait ainsi :

(1) ur ye-čči ara[1]

[1] La transcription est orthographique, sauf à l'intérieur des courbes obtenues par Praat, lorsque les caractères spéciaux n'étaient pas disponibles. Les abréviations sont les suivantes : SUJ= indice de personne sujet ; ACCUS= clitique accusatif ; DAT= clitique datif ; 1, 2, 3= personnes ; S= singulier ; P= pluriel ; PROX= particule proximale ; COP= copule ; IRR= particule d'irréel/potentiel ; CONC= particule de concomitance ; NEG= particule de négation ; POSTNEG= élément postnégatif ; NEGEX= négation d'existence ; REL.IRR= relateur irréel-potentiel ; ACC= accompli ; INACC= inaccompli ; ACCNEG= accompli négatif ; AOR= aoriste ; QLT= verbe de qualité ; ABS : cas absolu (état libre) ;

NEG SUJ3MS-manger.ACCNEG POSTNEG
Il n'a pas mangé.

Nous avons montré que seul *ur* était véritablement négatif, et que la présence de l'élément postverbal, loin d'être obligatoire, était fonction d'un ensemble de critères très précis, que nous avons étudiés dans Mettouchi (2000, 2001).
Les contextes d'absence de *ara* (48% des négations dans le corpus écrit[2] utilisé pour ce travail) sont les suivants :

- lorsqu'un indéfini est topicalisé,

(2) acemma ur t id i-qqar
chose NEG ACCUS3MS PROX SUJ3MS-dire.INACC
Il ne dira rien

- en contexte de serment ou d'énoncé catégorique,

(3) wellah ur t swi-γ
par-Dieu NEG ACCUS3MS boire.ACCNEG- SUJ1S
Je te jure que je ne l'ai pas bu !

- en coordination négative,

(4) ur uli-n yexxamen
NEG monter.ACCNEG- SUJ3MP maisons.INT
Il n'y a eu ni construction de maisons, ni

ur te-rbiḥ tfellaḥt
NEG SUJ3FS-prospérer.ACCNEG travail.agricole.INT
progrès dans la production agricole

INT= cas intégratif (état d'annexion) ; ANAPH= marqueur d'anaphore. Les affixes sont séparés par un trait d'union, les clitiques par un signe 'égale' ; les mots tronqués sont marqués par le signe '#'.
[2] Composé de plus de 300 formes verbales négatives.

- en subordonnée oppositive,

(5) la s slufu-γ ur faq-eγ
CONC DAT3S caresser.INACC-SUJ1S NEG se.rendre.compte.ACC-SUJ1S
Je la caressais sans m'en rendre compte

- en relative restrictive générique.

(6) ne-xdem ayen ur ne-ssin
SUJ1P-faire.ACC ce.que NEG SUJ1P-savoir.ACCNEG
Nous avons fait des choses qui nous étaient inconnues (litt. nous avons fait ce que nous ne savions pas faire)

En revanche, *ara* est systématiquement présent en subordonnée hypothétique :

(7) ma ur n-ruḥ ara s axxam
si NEG SUJ1P-aller.ACCPOSTNEG à maison.ABS
Si nous ne rentrons pas à la maison,

a γ čč-en lewḥuc
IRR DAT1S manger.AOR-SUJ3MP monstres.INT
les bêtes sauvages vont nous manger.

Il apparaît également dans les cas, fréquents, où le jugement négatif est de type constatif et collaboratif (en réponse à des questions par exemple) :

(8) ur t id y-ufi ara
NEG ACCUS3MS PROX SUJ3MS-trouver.ACCNEG POSTNEG
Il ne l'a pas trouvé.

Cette série de cas montre qu'en énoncé indépendant ou coordonné, la présence de *ara* est impossible lorsqu'un indéfini est thématisé, en contexte de serment ou d'énoncé catégorique, en coordination négative, en subordonnée oppositive, ou en relative restrictive générique.

Nous en tirons les conclusions suivantes[3] :

- En énoncé indépendant, sans *ara*, *ur* permet d'insister sur le résultat du parcours des situations envisagées : vide, ou absence. *Ara* quant à lui met en relief la situation de référence en tant qu'elle est en contradiction avec la situation envisagée. Si l'on transpose ceci dans la problématique de la co-énonciation, on voit que *ur* met en place un jugement négatif absolu, égocentré, qui ne repose pas sur un terrain commun entre énonciateur et co-énonciateur. Avec *ur*, nous sommes dans l'altérité radicale, qualitative. Avec *ara*, le jugement négatif n'est plus égocentré, il repose sur un terrain partagé avec le co-énonciateur.

- En énoncé coordonné, *ara* n'est pas possible car il stabilise chaque jugement négatif par rapport à la situation d'énonciation, qui comprend le co-énonciateur. Or la coordination nécessite que chaque jugement négatif soit repéré par rapport à celui qui précède et/ou celui qui suit. Le lien est syntagmatique et c'est l'ensemble des propositions coordonnées qui est repéré par rapport à la situation d'énonciation.

- En énoncé subordonné, nous sommes dans le domaine de la préconstruction. En effet, une subordonnée a pour caractéristique de n'être pas directement repérée par rapport à la situation d'énonciation, contrairement à la principale. Ceci implique que la co-énonciation soit moins importante ici que la référenciation. Il s'avère qu'*ara* est régulièrement associé aux cas marquant un ancrage référentiel par opposition aux cas où le qualitatif, l'attribution de propriété priment.

A partir de ces contraintes d'emploi, nous avons pu assigner à *ara* une valeur liée à son origine étymologique ('chose'), et considérer que ce marqueur permettait d'attribuer au jugement négatif une certaine stabilité, en l'ancrant référentiellement, et dans la situation d'énonciation.

2. Analyse intonative

Voyons à présent s'il existe des profils intonatifs typiques des différents cas de figure morphosyntaxiques et énonciatifs dégagés en 1[4].

[3] Pour plus de détails voir Mettouchi (2000) et (2001).
[4] Les analyses acoustiques sont menées sur le logiciel Praat, www.fon.hum.uva.nl/**praat**/

Les données présentées dans cet article proviennent d'un travail de terrain sur une variété kabyle centrale (tribu des Aït Idjer, village d'Aït Ikhlef, commune de Bouzeguène). Les exemples sont extraits de deux contes traditionnels racontés de manière improvisée par madame Tousia Rabia[5]. Nous sommes donc en contexte de monologue, dirigé vers un auditoire.

Dans le premier conte, un veuf vit avec ses sept filles. Une voisine souhaiterait l'épouser. Elle se fait bien voir de ses filles qui réclament au père qu'il épouse cette femme. Il le fait, mais celle-ci exige alors qu'il abandonne ses filles. Une première fois, le père les abandonne lors d'une corvée de bois dans la montagne. Grâce à la petite Yamina, les filles retrouvent le chemin de la maison. A la demande de la marâtre, le père creuse une fosse et jette ses filles dedans. La petite Yamina jette un sort à son père. Grâce à Yamina, les petites filles sont délivrées et prennent possession de la maison d'un chat sauvage, après l'avoir tué. Le père retrouve ses filles, ils se pardonnent mutuellement, et les filles jouent un mauvais tour à la marâtre pour la punir.

Dans le deuxième conte, un veuf vit avec ses sept filles, qu'il doit laisser seules pour aller en pèlerinage à la Mecque. Il leur recommande de se méfier de tous et de n'ouvrir à personne. Mais une ogresse convainc l'aînée de lui ouvrir, et dévore six des sept soeurs. La benjamine parvient à s'échapper et à se réfugier chez un couple de vieillards, mais un sultan la marie de force à son fils. La jeune fille devient muette, et ne recouvre la parole qu'en retrouvant son père, qui la cherchait à travers le pays déguisé en mendiant.

2.1. Enoncés simples

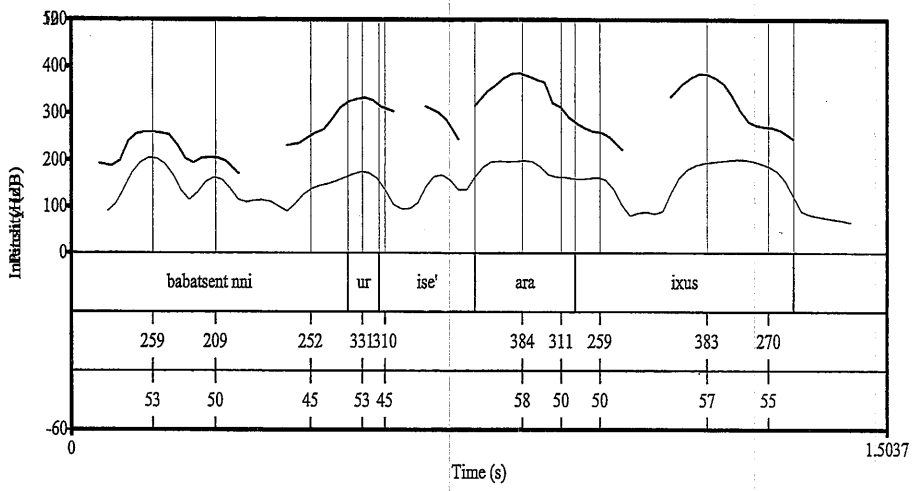
Nous considérons comme tels les énoncés qui ne comportent ni relateur ni complémenteur, et qui présentent un contour intonatif montant descendant délimitant l'ensemble de la structure. Notre premier exemple (sur un corpus de 37 séquences analysées pour les deux contes, pour un

[5] Le texte complet se trouve dans Mettouchi (2006a) et (2006b). L'enregistrement peut être écouté sur le site du CRDO (Centre de ressources pour la Description de l'Oral), http://crdo.risc.cnrs.fr/exist/crdo/.

total de 19 mn en débit normal) est extrait du passage où la conteuse indique que le père a emprunté des robes à ses voisins pour en vêtir ses filles car il était pauvre et n'avait pas les moyens de les acheter.

(9) babatsent-nni ur i-sɛi ara i-xus□ //
père_leur-ANAPH NEG SUJ3MS- posséder. ACCNEG POSTNEG SUJ3MS - manquer. ACC //
leur père ne possédait pas d'argent il était pauvre //

Courbe 1 Conte 1 exemple (9)[6]



Sur le plan intonatif, cet énoncé présente le profil caractéristique des structures comportant le marqueur *ara*. Un premier pic apparaît à 331 Hz sur *ur*, le second à 384 Hz sur la première voyelle de *ara*. Le différentiel entre *ur* et la syllabe suivante est de 53 Hz.
Tous les énoncés du corpus comportant *ara* présentent cette montée caractéristique sur le marqueur postverbal.

[6] La première ligne contient la transcription, la seconde les valeurs de F0 en Hertz, la troisième celles de l'intensité, en décibels. L'échelle est en général de 0 à 500 Hz, et de moins 60 à +120 dB. La courbe de F0 est en trait plein, celle de l'intensité en discontinu.

Dans l'exemple suivant, qui ne comporte pas *ara*, la conteuse raconte ce qui est arrivé à la jeune fille après avoir été épousée contre son gré et sans l'agrément de son père.

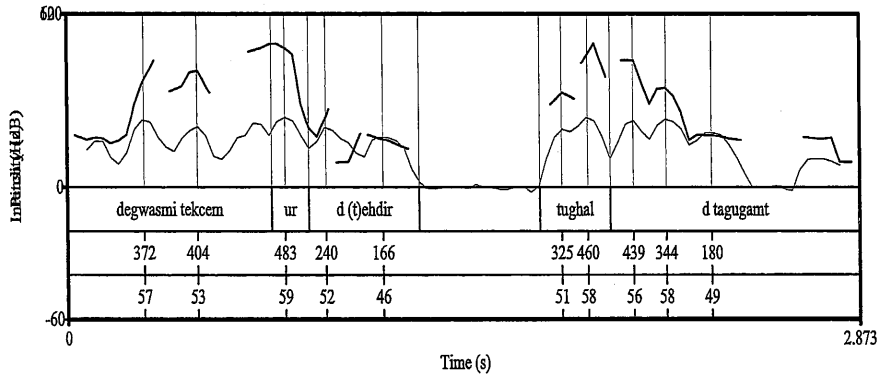
(10) taqcict tamcumt meskint
fille.ABS malchanceuse pauvre
La pauvre malheureuse

degwasmi te-kcem ur d=te-hdiṛ //
depuis SUJ3FS-entrer.ACC NEG PROX=SUJ3FS-parler.ACCNEG
depuis qu'elle avait pénétré sa nouvelle maison (en tant qu'épouse) elle ne disait mot.

t-uγal d tagugamt //
SUJ3FS-devenir.ACC COP muette.ABS
elle était devenue muette

L'énoncé négatif est précédé d'une subordonnée adverbiale de temps qui situe le jugement négatif par rapport à une période. Le pic de F0 est sur *ur*, sa valeur est de 483 Hz, pour une valeur de 180 Hz en fin d'énoncé et de 372 à 404 Hz sur la partie adverbiale.

Courbe 2 Conte 2 exemple (10)



Cette fois, *ur* est le sommet mélodique de l'énoncé, et le différentiel entre ce marqueur et la première syllabe du prédicat est de 243 Hz. Ce différentiel est considérable et joue sur la perception de *ur* comme pic marqué de F0.

C'est la même personne qui raconte les deux contes, et l'on peut voir qu'ici *ur* est également, en valeur absolue, plus élevé que dans les énoncés comportant *ara*. Les autres énoncés négatifs simples de notre corpus présentent des profils similaires : une désaccentuation importante du prédicat après *ur*, lui-même caractérisé par un pic de F0.

L'exemple suivant est extrait de l'explication de la petite fille à l'ogresse : elle lui rapporte les recommandations du père: partant en pèlerinage, il rappelle à ses filles que personne n'est censé se présenter à la maison pour demander à entrer, puisqu'elles n'ont aucune famille.

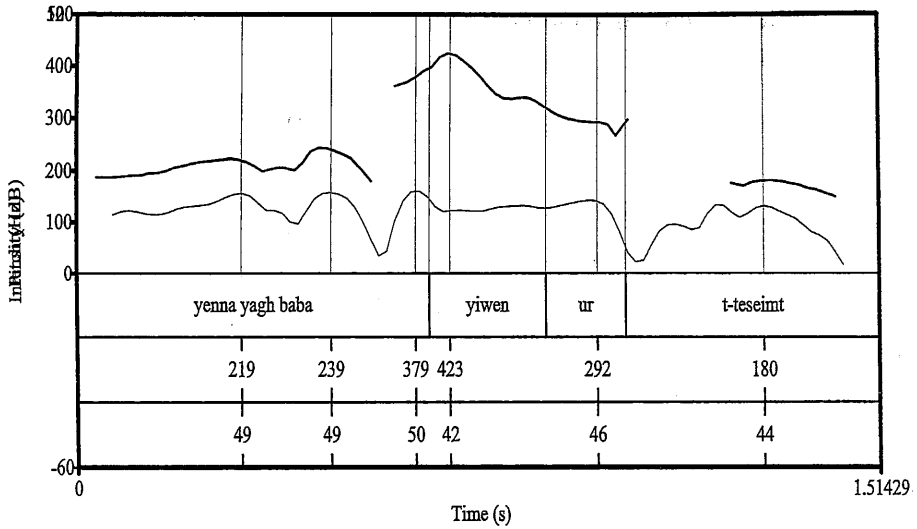
(11) nna-nt=as nekkenti ur ne-sɛi ara
dire.ACC-SUJ3FP=DAT3S nous.F NEG SUJ1P-posséder.ACCNEG POSTNI
elles lui dirent nous nous n'avons pas

xwal# xal-nteγ /
oncles# tantes-nos
d'oncles de tantes

ye-nna=yaγ baba
SUJ3MS-dire.ACC=DAT1P papa
notre père nous a dit

yiwen ur t=te-sɛi-mt //
un NEG ACCUS3MS=SUJ2FP-posséder.ACCNEG-SUJ2FP
'Vous n'avez personne (au monde)'

Courbe 3 Conte 2 Exemple (11)



On remarque que là aussi, le différentiel entre la valeur de *ur* (292 Hz) et celle de la voyelle /i/ (180 Hz) du prédicat : il est de 112 Hz. L'énoncé comportant une topicalisation d'indéfini, c'est cet indéfini, *yiwen*, qui porte le pic de F0 à 423 Hz, *ur* ne constituant qu'un pic secondaire, néanmoins nettement plus élevé que les autres proéminences de l'énoncé.

L'étude des énoncés simples nous permet de dégager deux profils intonatifs principaux :

1- L'un, sur les énoncés comportant *ara*, se caractérise par un premier pic de hauteur variable sur *ur*, et un deuxième, toujours plus haut que le premier, sur *ara*, qui est l'élément focalisé de l'énoncé. Plus la F0 sur *ara* est élevée, plus l'énoncé est polémique.

2- L'autre, sur les énoncés ne comportant que *ur*, se caractérise par un pic de F0 sur *ur*, suivi d'une désaccentuation ou d'une chute rapide de

la F0 sur le reste du prédicat. Plus le prédicat est désaccentué et/ou plus le pic sur *ur* est élevé, plus la négation est absolue ou hyperbolique.

Sur le plan iconique, les montées de F0 signalent que le contenu prédicatif est en débat, il y a appel à l'autre, nous sommes dans la co-énonciation. La position des pics distingue le cas où le débat concerne un jugement négatif ancré en situation, objet d'une négociation (pic principal sur *ara*), de celui où le jugement négatif est lié à un parcours de situations, et présenté comme hors situation (pic principal sur *ur*, le marqueur *ara* n'étant pas présent dans l'énoncé).

Le fait qu'un pic soit toujours présent sur *ur* est d'ordre typologique : en kabyle, les particules (particules aspecto-modales, négation, relateurs) sont en tête de proposition, et sont systématiquement accentuées. Ce pic est plus marqué lorsqu'*ur* n'est pas associé à *ara*. Comparativement, il perd un peu de sa saillance lorsqu'*ara* est présent dans l'énoncé, puisque c'est ce dernier qui porte alors le pic principal de F0.

L'intonation caractéristique des négations comportant *ara* est très proche de celle des énoncés dont le complément direct lexical est un indéfini, comme dans (12) :

(12)	ur	ye-swi	tiqit
	NEG	SUJ3MS-boire.ACCNEG	goutte.ABS

Il n'a rien bu

Ceci nous amène à considérer qu'en liaison avec sa grammaticalisation avancée (perte de la valeur sémantique de 'chose'), *ara* fonctionne comme un indéfini, marqueur de quantité minimale. Il ne fonctionne pas prosodiquement comme un nominal référentiel car il n'a pas l'intonation caractéristique des compléments de verbes transitifs.

Voyons à présent comment se comportent les énoncés complexes négatifs sur le plan prosodique.

2.2. Enoncés complexes

Nous avons considéré comme tels les énoncés comportant plusieurs propositions reliées entre elles, soit par coordination, soit par subordination, étant entendu qu'il est fréquent en kabyle que les relations

de dépendance ne soient pas marquées segmentalement, ce qui implique que la prosodie ait un rôle important à jouer.

Dans l'exemple suivant, le père explique à ses filles qu'elles n'ont aucune famille, et par conséquent, qu'elles ne devront ouvrir la porte à personne en son absence.

(13) ye-nna=yas ur te-sɛi-mt babatkwent
SUJ3MS-dire.ACC=DAT3S NEG posséder.ACCNEG-SUJ2FP SUJ2FP-père.votre
Il leur dit 'vous n'avez pas de père

aql=iyi ulac=iyi /
voici=ACCUS1S NEGEX=ACCUS1S
me voici je ne suis plus là

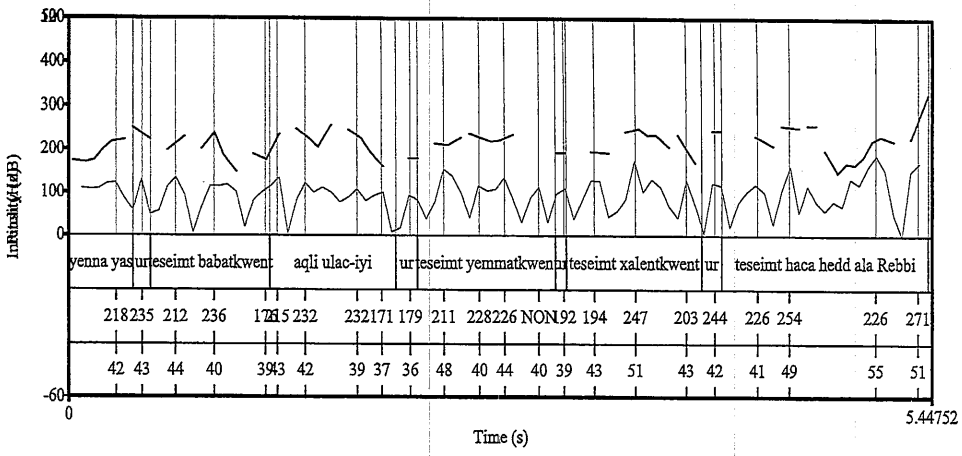
ur te-sɛi-mt yemmatkent /
NEG SUJ2FP-posséder.ACCNEG-SUJ2FP mère.votre
vous n'avez pas de mère

ur te-sɛi-mt xalnkent /
NEG SUJ2FP-posséder.ACCNEG-SUJ2FP tantes.votre
vous n'avez pas de tantes

ur te-sɛi-mt ḥaca ḥedd
NEG SUJ2FP-posséder.ACCNEG-SUJ2FP seulement personne
vous n'avez personne au monde

ala Rebbi //
seulement Dieu
seulement Dieu'

Courbe 4 Conte 2 exemple
(13)



On peut remarquer un mouvement crescendo dans les contours intonatifs, ainsi qu'au niveau des pics sur *ur* composant cette suite de jugements négatifs coordonnés. Aucun morphème ne marque la coordination sur le plan segmental, c'est l'absence de *ara* qui, en ne situant pas les propositions indépendamment les unes des autres, induit la liaison. La première partie de la séquence comporte un énoncé qui pourrait être interprété comme simple (hors de la séquence coordonnée), mais l'identité de contour prosodique avec les autres propositions négatives montre qu'il en est le premier membre, et que c'est une incise qui le sépare des trois autres propositions.

Chaque *ur* est intoné avec un différentiel (resetting) de 8 à 40 Hz par rapport à la dernière syllabe de la proposition précédente (de 218 à 235 Hz, de 171 à 179 Hz, de 203 à 244 Hz)[7]. La valeur du pic de F0 sur *ur*

[7] Le différentiel entre la deuxième et la troisième proposition négative est impossible à calculer car les consonnes de fin de proposition ne sont pas voisées, ce qui empêche d'avoir des valeurs de F0. De même, les valeurs de F0 à l'intérieur du premier intervalles ne sont pas calculées car il y a une erreur dans la courbe sur le relateur *ara*.

augmente d'une proposition à l'autre, de même que les valeurs de F0 à la fin de chaque proposition. C'est ce qui permet l'interprétation comme suite coordonnée : la coordination implique un repérage en chaîne des sommets, et une répétition de contours semblables.

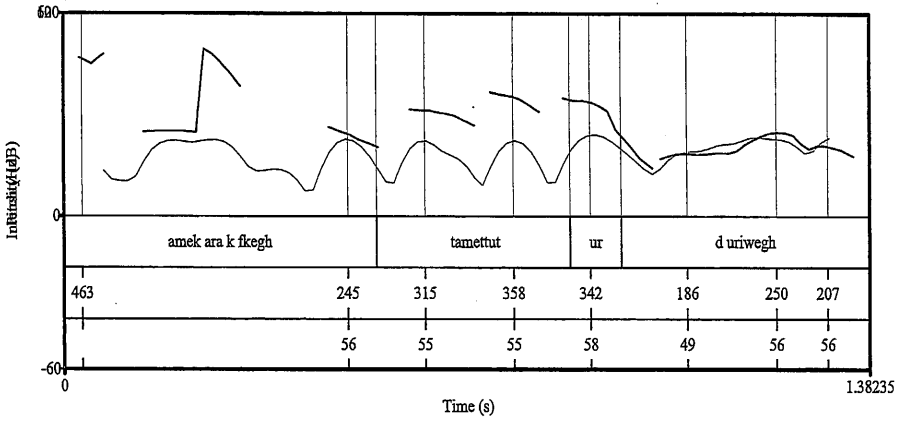
L'exemple suivant est différent en ce qu'il contient un antécédent, *tameṭṭut*, et une relative négative restrictive, indispensable à la construction de la référence de cet antécédent. Il s'agit de la réponse au sultan du vieil homme chez qui la jeune fille s'est réfugiée.

(14) amek ara k=fke-γ
comment REL.IRR ACCUS2MS=donner.AOR-SUJ1S
comment pourrais-je te donner

tameṭṭut ur d=uriw-eγ
femme.ABS NEG PROX=donner.naissance.ACCNEG-SUJ1S
une femme à laquelle je n'ai pas donné naissance

ad=tt t-aγ-eḍ //
IRR=ACCUS3FS SUJ2S-prendre.AOR-SUJ2S
pour que tu l'épouses ?

Courbe 5 Conte 2 Exemple (14)



La F0 est élevée sur l'antécédent (315 puis 358), puis descend légèrement sur *ur* (342 Hz). Ensuite elle diminue très fortement pour atteindre 185 Hz, puis 227 Hz et 203 Hz sur les trois syllabes du prédicat.

La subordonnée relative restrictive, par son profil intonatif, fait bloc avec l'antécédent : *ur* se situe environ à la même hauteur que l'antécédent, puis le prédicat de la relative, comme celui des principales, est désaccentué. Avec *ara* la relative ne serait plus restrictive, et prosodiquement ce serait *ara* qui serait le sommet mélodique de l'énoncé. On remarque que bien que la F0 ne soit pas à son point le plus élevé sur *ur*, l'intensité est maximale. Ceci compense perceptivement la baisse de F0.

L'exemple suivant est composé d'une subordonnée hypothétique et d'une principale, avec dans chacune une négation comportant le marqueur postverbal *ara*. Il s'agit de la condition mise par le père à son remariage : il souhaite que la plus petite de ses filles soit assez grande pour qu'elle ne souffre pas de sa marâtre si jamais celle-ci s'avérait mauvaise.

(15) ma ur te-ttawed□ ara
si NEG SUJ3FS-atteindre.INACC POSTNEG
si elle ne parvient pas

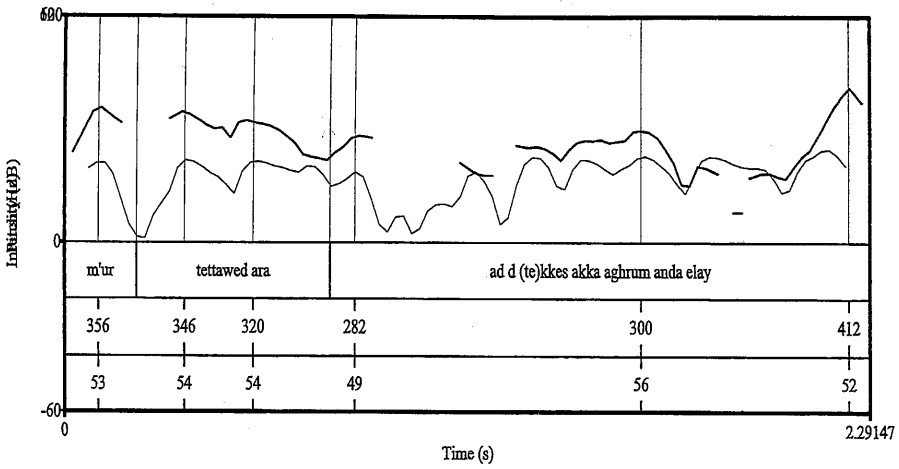
ad=d te-kkes akka aγrum
IRR=PROX SUJ3FS-ôter.AOR ainsi pain.ABS
à ôter ainsi le pain

anda ɛlay /
où être.haut.ACC.QLT
de cet endroit élevé

ur qebbl-eγ ara
NEG accepter.INACC-SUJ1S POSTNEG
je n'accepterai pas

ad zewǧ-eγ //
IRR se.marier.AOR-SUJ1S de me marier

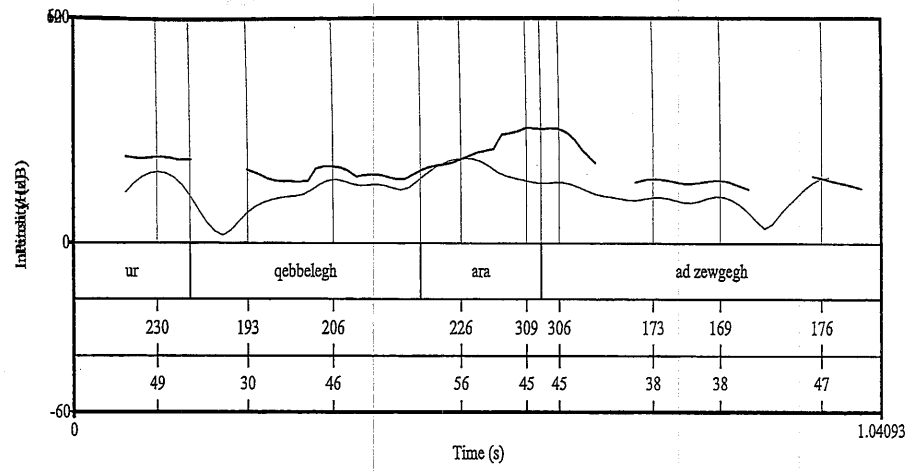
Courbe 6 Conte 1 Exemple (15)



Dans la protase de cet énoncé complexe hypothétique, *ur* est cliticisé sur le marqueur d'hypothèse (*ma+ur* = *m'ur*), et intoné en plage haute, c'est un sommet de F0 (356). *Ara* est intoné plus bas, à 320 Hz: il n'y a pas débat, mais mise en place d'une proposition préconstruite, qui sert de point d'ancrage à une apodose. La remontée finale sur *ɛlay* est continuative, elle annonce l'apodose.

Dans l'apodose, *ur* est intoné en plage médiane, tandis que *ara* connaît un pic de F0 à 309 Hz sur la deuxième syllabe, sans doute provoqué par l'assimilation avec la particule *ad*, tête de la complétive. La première syllabe de *ara* est à 226 Hz, mais l'intensité est à son maximum, permettant qu'*ara* soit perçu comme saillant.

Courbe 7 Conte 1 Exemple (15)



Dans la protase le pic sur *ur* permet d'attirer l'attention sur le caractère négatif de la condition posée ; dans l'apodose, *ur* est intoné plus bas (230 Hz) : une baisse régulière de l'intonation permet de marquer la dépendance de l'apodose par rapport à la protase. Le schéma des hypothétiques place *ara* plus bas que *ur*.

Conclusion

Ur est toujours intoné en plage haute, et ceci est lié à son statut de particule fonctionnant comme tête syntaxique. En énoncé simple, il est le sommet de l'énoncé lorsqu'il n'est pas associé à *ara*. En énoncé complexe, *ur* n'est pas nécessairement le sommet principal, mais il est proéminent, en général à un niveau plus bas que la proposition précédente lorsqu'il y a subordination, plus haut lorsqu'il y a coordination.

Lorsqu'un énoncé simple comporte *ara*, ce dernier est le sommet de l'énoncé. Lorsqu'*ara* apparaît en énoncé complexe, il est en général intoné plus bas que *ur*.

La co-énonciation joue un rôle important dans la prosodie des énoncés simples, tandis que dans les énoncés complexes, le rôle principal de la prosodie est de structurer les relations syntaxiques.

Parmi les nombreux points laissés en suspens par cette étude exploratoire, la question du rôle de la prosodie dans les repérages interpropositionnels nous paraît centrale : en effet, les quelques tendances dégagées pour la négation en énoncé complexe sont sans doute à replacer dans un cadre plus large, celui du marquage des relations de dépendance syntaxique.

BIBLIOGRAPHIE

HIRST, D. et DI CRISTO, A. (eds) (1998), *Intonation Systems: A Survey of Twenty Languages*. Cambridge University Press, Cambridge.

MAGRO, E.-P. (2003) On Some Intonational Features of French and Maltese spontaneous Speech : a Comparative Study, Proceedings of CIL17, Prague, CD-Rom.

METTOUCHI, A. (2000) « Négation, co-énonciation et référenciation : le marqueur *ara* en kabyle de l'Ouest », dans *Comptes-Rendus du GLECS tome XXXIII (1995-1998)*, Publications des LANGUES'O, Paris, 87-104.

METTOUCHI, A. (2001) « La grammaticalisation de *ara* en kabyle, négation et subordination relative », dans *Travaux du CerLiCO n°14*, Col G. et Roulland D. (eds), P.U.Rennes, 215-235.

METTOUCHI, A. (2006a) « Un conte kabyle » in *Studi Berberi E Mediterranei, Miscellanea offerta in onore di Luigi Serra*, A.-M. di Tolla (ed), « Studi Magrebini » vol. IV NS, Napoli, Università degli Studi di Napoli « L'Orientale », 105-120.

METTOUCHI, A. (2006b) – « Anaphoricité et appel à l'attention partagée dans un conte oral en kabyle (berbère) » in *Loquentes Linguis, Studi linguistici e orientali in onore di Fabrizio A. Pennacchietti*, P.-G. Borbone, A. Mengozzi & M. Tosco (eds), Wiesbaden: Harrassowicz, 499-507.

MOREL, M.A. & DANON-BOILEAU, L. (1998) - *Grammaire de l'intonation : l'exemple du français*, Paris, Ophrys.

WICHMANN, Anne (2000) - *Intonation in Text and Discourse : Beginnings, middles and ends*. Harlow: Pearson and Longman.

From Fieldwork to Annotated Corpora : The CorpAfroAs project

Amina Mettouchi* & Christian Chanard**

INTRODUCTION

In the first years of this new century, in the domain of linguistics, much emphasis is being put on language diversity, as well as on language technologies. Not so long ago, grammatical theories were content to rely on a small number of well-described European and Asian languages, and corpora-design was limited to some of those well-known languages[1]. With the development of typology, and the growing concern about the fast disappearance of hundreds of the estimated 6000 languages currently spoken on our planet, language descriptions are now given more and more importance. In the meantime, language technologies have become more and more accessible to the linguist, through the generalization of the use of computers, and the availability of high-quality portable recording devices. The first result of this technological revolution was the development of language archives aiming at preserving the work of fieldwork linguists through the digitalization of recordings and transcripts. Such initiatives as the LACITO Archive[2], the CRDO[3], or DOBES[4] or other centers for the preservation of language diversity and endangered languages have emerged. A number of texts in a great variety of languages have thus been digitalized. However, annotations are not always provided, and when they are, they are not standardized and/or do not allow complex queries in the database. CorpAfroAs[5], a project funded by the Agence Nationale de la Recherche (ANR) in France, has emerged in this context, as a pilot corpus aiming at providing a structured database of spontaneous

* EPHE, Paris.
** CNRS LLACAN, Villejuif.

[1] See for instance such initiatives as the London-Lund Corpus of spoken English, the British National Corpus, C-Oral Rom, etc.

[2] http://lacito.vjf.cnrs.fr/archivage/presentation.htm

[3] http://crdo.risc.cnrs.fr/exist/crdo/

[4] http://www.mpi.nl/DOBES

[5] The CorpAfroAs project is conducted by three French research laboratories, and associate French and International researchers. The principal coordinator is A. Mettouchi, the associate coordinators are M. Vanhove and D. Caubet. Two experts are following the project and providing feedback and guidelines: B. Comrie (MPI Leipzig and UCSB), and S. Izre'el (University of Tel-Aviv). The complete list of members can be found on http://web.me.com/aminamettouchi/CORPAFROAS/ Abstract.html.

recordings of Afroasiatic languages, transcribed, translated, and annotated in view of allowing complex queries.

The ultimate goal of CorpAfroAs is to trigger a number of similar endeavors for various language families. This is why the design of the corpus, and the scientific decisions made, must be brought to the knowledge of the community, and proposed for discussion and implementation. Hence this paper, which is an update on a preceding paper (Mettouchi & al. 2010) presenting the main lines and goals of CorpAfroAs.

Our aim here is to focus on the theoretical and technical developments of the project. In part 1, we present the motivations for the choice of software made for the project, in part 2, we focus on the design of annotation tiers in relation to some queries relevant for our language family, and in part 3 we briefly present our metadata form.

1. General design and annotation procedure

CorpAfroAs is organized along two axes, linked to the nature of the materials and to the aim of the project, which is typological comparability among languages: prosodic analysis, and morphosyntactic glossing.

The body of data is spoken, and we have decided to fully take into account this oral dimension by working on segmentation. We do not use the punctuation system of written texts, because it is not adapted to the specific nature of the spoken language (Wichmann 2000). Instead, we are adapting the system of boundary-marking used for instance in the C-ORAL-Rom developed by Cresti & Moneglia[6].

We therefore analyze the prosodic units of our languages into minor (non-terminal) and major (terminal) units, using the software Praat[7]. No other specification (tones, contours, etc.) is given to those boundaries, but the fact that the transcription is indexed to the sound, itself available in .wav format, will allow more in-depth prosodic studies on the available data.

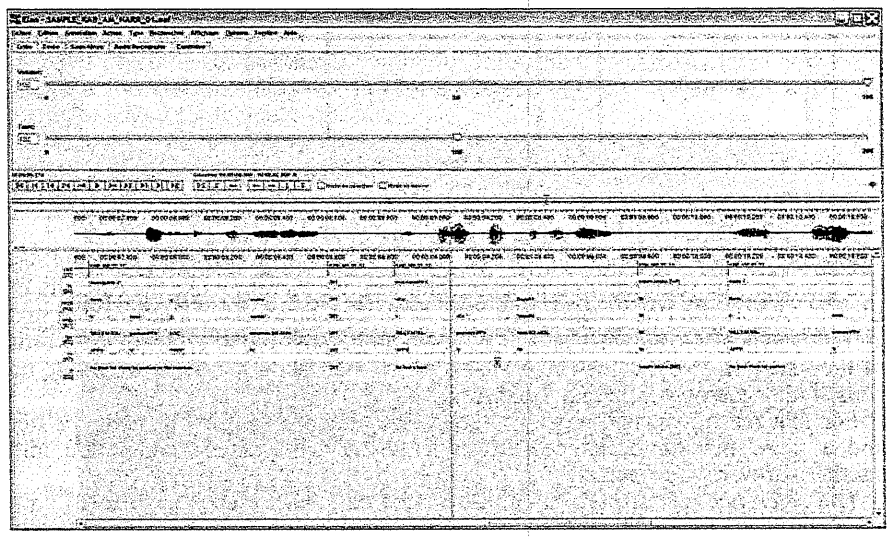
Segmentation by native speakers provides the basis for analysis of the major (terminal) intonation-units, which turn out to be based on cues used in a wide variety of languages, namely pitch reset, lengthening, anacrusis, and pauses. Minor intonation unit are typically more difficult to define. The fact that the sound file will be linked to the segmented transcription will facilitate alternative proposals by other researchers.

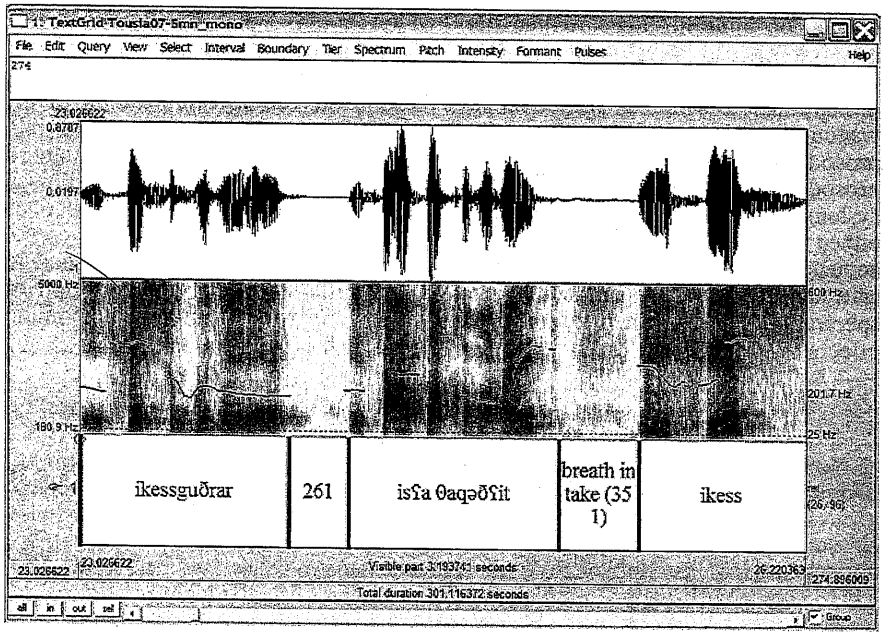
The software in which CorpAfroAs is designed and will ultimately be put online is ELAN[8], developed by the Max Planck Institute in Nijmegen. This software was chosen for a number of reasons: it is dedicated to the creation of complex annotations on video and audio resources, it is open-source and free, it is regularly updated, and it is widely used among linguists. Annotations are

[6] http://lablita.dit.unifi.it/coralrom/
[7] Paul Boersma & David Weenink, http://www.fon.hum.uva.nl/praat/
[8] http://www.lat-mpi.eu/tools/elan/

created on multiple layers, called tiers, which can be hierarchically connected, and time-aligned to the media.

In Elan, the sound is only available through the online visualization of the waveform. Although it is theoretically possible to segment the sound into prosodic units using this visualization, the results are actually too inaccurate: a prosodic segmentation made into Elan and checked under Praat showed systematic misalignment of the segments: the boundary was a few milliseconds either to the right or to the left of where it should have been marked, given the acoustic information provided by Praat. This is the reason why we decided to start the segmentation process in Praat. This software allows visualization of spectrum, pitch and intensity, on top of the waveform, with important zooming effects.





However, Praat has no hierarchical structure allowing complex annotations and queries. The segmentation process achieved, we therefore use the capacity of Elan to import the Praat segmented file. At this point we have a segmented text correctly synchronized with the sound. It would be possible to segment the words of the text into morphemes and annotate them into Elan, but there would be no consistency guaranteed in this hard work.

For that reason we just prepare the text into Elan, by adding a *reference* tier and a *word* tier for each speaker, to allow morphosyntactic annotation into another software, Toolbox.

The *reference* tier displays a unique numbered label for each segmented unit, to identify it for later referencing. This labelization can be automatically generated by Elan.

The *word* tier contains each word of the text tier in a separated cell. It can be automatically generated by Elan text tokenizer, provided the text in the *tx* tier is transcribed without sandhis, and normalized to some extent. If not, that is if we choose to have a *tx* tier transcribed with assimilations, into phonetic words that cannot simply be segmented into smaller units to form grammatical words, then an intermediary line is necessary, with a transcription into underlying forms, separated by spaces, that can then be tokenized into words (the 'mot' tier).

When the text is segmented into words, we can export it into Toolbox[9]. Toolbox is a software dedicated to the management of textual databases such as lexicon and/or phrase databases. In addition, it can annotate a text with the contents of a lexicon. This is an interactive process in which the software searches the lexicon for each word of the text to interlinearize, and proposes the glosses it finds, each one on a line, vertically aligned under the word. If the actual word doesn't exist in the lexicon, Toolbox tries to isolate possible affixes (which may be listed in the same lexicon or in a special one), glosses them if they exist, until it finds the root in the lexicon, or, if not, outputs a failure mark for the rest of the word.

The user has to interactively choose between gloss ambiguities, correct wrong segmentation or add new lexemes into the lexicon with their glosses. This ensures a better level of consistency in the morphosyntactic annotation process.

2. The tiers in ELAN and their technical and theoretical motivations

After much discussion and a number of tests, the CorpAfroAs team decided to adopt a format containing six linguistic annotation tiers. As we will see, other tiers are added for technical reasons.

2.1. The technical organizing principles

The **ref** line references each segment by a numbered label. It is the only one which is synchronized to time, the *tx* tier being in *symbolic association* to it, that is to say they share the same time segmentation. This *ref* tier is the ultimate reference that subsumes the other tiers. So, any cell, in the end, refers to a *ref* segment parent, and this allows Elan to jump to that main segment when asked to.

tx: is the tier in which the text is transcribed in broad phonetics, into phonetic words (with assimilations, sandhis, etc.). Major and minor boundaries are indicated (/ & //), and pauses over 200 ms appear in a separate unit.

mot: is the tier in which the text is transcribed into grammatical words, with no morphemic separators (- =), and using a phonological (i.e. 'regularized' as compared to the broad phonetics one) transcription.

mb: is the tier in which the text is segmented into morphemes (one cell per morpheme); - goes in the cell that contains the affix, = goes in the cell that contains the clitic.

ge: is the tier in which a gloss is provided for each morpheme cell. The glossing is into grammatical category labels, and is based on the Leipzig

[9] http://www.sil.org/computing/toolbox/

Glossing Rules[10]. Other relevant information (parts of speech, verb class, syncretism phenomena, etc.) goes into tier **rx**.

rx: is the tier in which all information relevant and necessary for retrieval purposes is entered. If there is more than one label per cell, we separate them with a slash.

ft: is the tier where the text is translated (free translation). The translation is indexed to minor or major units depending on the syntax of the language.

The principle of Elan is to document a media resource (audio or video signal). The signal is displayed on a horizontal timeline, and tiers can be created under that line to synchronically annotate the signal: there is a vertical correspondence between the annotation lines and the signal line.

All the annotation lines do not need to be directly synchronized to the signal. In the CorpAfroAs project, only the first tier (ref) corresponding to the segmentation into prosodic units of *tx* is synchonized. The other tiers are indirectly indexed to time by dependency relationship among them.

The *mot* tier has a *symbolic subdivision* dependency with the *tx* tier. This means that the time duration of a text segment unit is divided (equally[11]), at the *mot* tier level, between the words that belong to that segment. These words are not synchronized to sound, but they share the same time segment than the text segment which they belong to. They lie within the time boundaries of their parent segment.

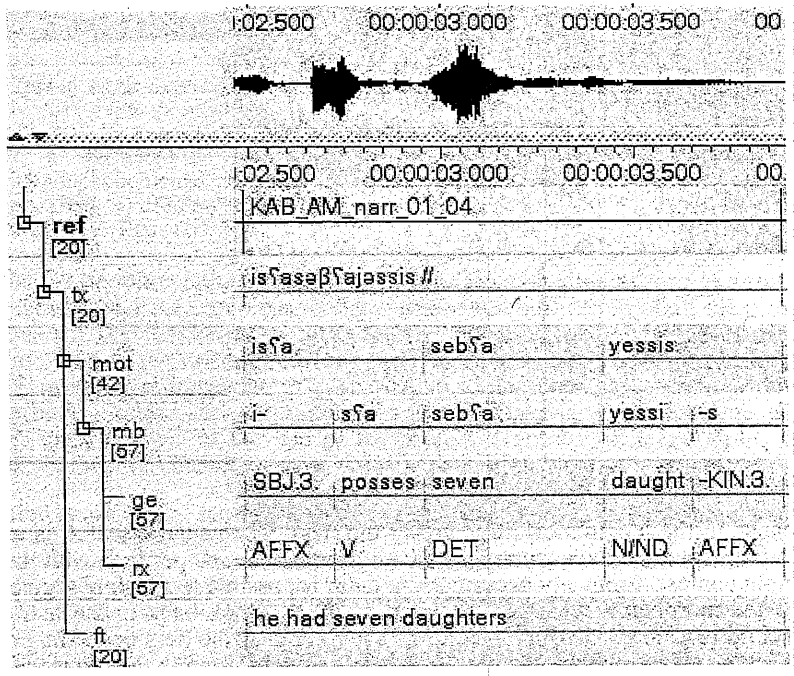
The *mb* tier is a *symbolic subdivision* of the *mot* tier, i.e. the different morphemes of a word share the time segment of the word they belong to.

The *ge* and *rx* tiers have a *symbolic association* dependency with the *mb* tier. That means there is a term to term correspondance between morpheme and gloss in *ge* and *rx*.

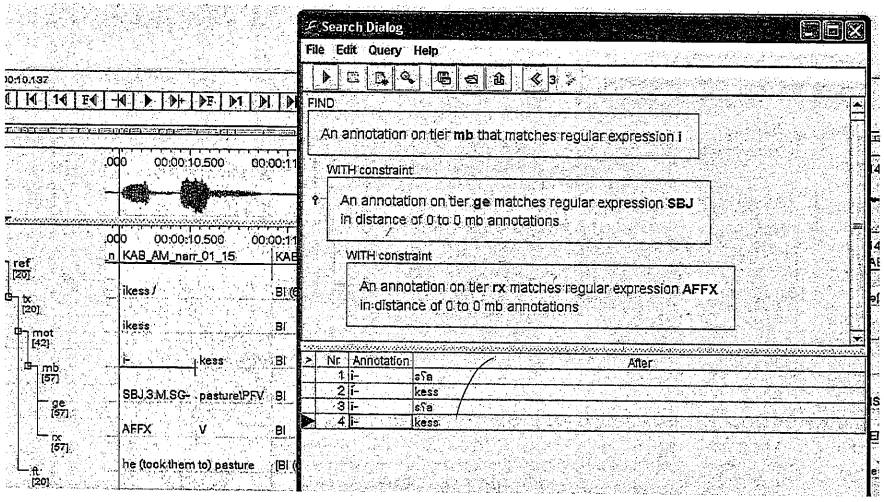
These dependencies from tier to tier - also called child and parent relations - makes it possible to vertically align the elements in the way linguists usually present interlinear texts.

[10] http://www.eva.mpg.de/lingua/tools-at-lingboard/glossing_rules.php

[11] The space occupied by a word with regard to its parent text segment does not represent its actual duration in the signal. This is why these are called *symbolic* subdivisions in Elan.



Elan has a retrieval engine allowing to look for a sequence (a word, a morpheme...) in a specific tier, with possible additional constraints (another sequence in another tier). The correspondence between the two (or more) tiers may be just direct, i.e. sharing the same time segment, or the second one may be searched within a certain distance from the segment of the first. In the example below, we are looking for all the morphemes (*mb* tier) 'i' that have 'SBJ' in their direct (distance 0) corresponding glossing tier (*ge*) and that are affixes ('AFFX' in the *rx* tier).



Thanks to the hierarchical structure of the tiers, when such a search is made, Elan will display all the occurrences of this morpheme, one per line, with the left and right context. From any occurrence, a jump is possible to the time segment to which the morpheme belongs, since Elan is able to look back in the hierarchy from child to parent. This time segment will display all the tiers depending on it, and clicking on the play button will make it possible to listen to the sequence.

In the same way, a concordance can be made for a sequence (morpheme, word, gloss…), which will display occurrences centered in the line, with the left and right context within a selected distance. Statistics can also be displayed.

2.2. The theoretical organizing principles

The **tx** line is the one that holds the transcription of prosodic units. Its purpose is to reflect as closely as possible the sound file, including false starts and other phenomena found in spontaneous speech. As the phonology of the language is known, the transcription is not completely phonetic, although it includes word-boundary phenomena (sandhi, etc.), as those may be interesting for the phonology-syntax interface.

The **mot** line is mainly an intermediary tier that allows the subsequent segmentation into morphemes. It contains grammatical words, the definition of those words being language-dependent, therefore, this tier may not reflect exactly the word segmentation of the *tx* tier. The **mb** line is segmented into morphemes, allowing for allomorphs and all such variation desirable for a varied morpheme inventory.

The **ge** line is the morpheme-by-morpheme gloss of the *mb* line. Its syntax is based on the Leipzig Glossing Rules:

> When a single object-language element is rendered by several metalanguage elements (category labels), these are separated by periods. Ex: 3.M.SG (Rule 4)
>
> When a single object-language element is rendered by several metalanguage elements (words), these are separated by underscores. Ex: be_tall (Rule 4A)
>
> If a grammatical property in the object-language is signalled by a morphophonological alternation (ablaut, mutation, tone alternation, etc.), the backslash is used to separate the category label and the rest of the gloss. Ex: write\PFV (Rule 4D)

The list of abbreviations provided by the LGR is incomplete, and therefore one of the tasks we have completed is the creation and unification of all the proposed glosses for the languages of our pilot-corpus, with the assistance of Bernard Comrie, one of the creators of the LGR. A number of problems arose, to which solutions were proposed. Those solutions have been implemented within the Afroasiatic phylum, but are exportable to other language families, and will be listed and published at the end of the project. Here are some examples of the issues that were discussed:

> Traditional labels: for each language family of the phylum, descriptive traditions going back sometimes to more than a century ago, have consecrated the use of some labels, such as 'suffixal conjugation' in Arabic, 'free state' in Berber. Those labels, although they have their motivation and are grounded in decades of analyses, make little or no sense to linguists that do not work within those traditions. We have decided to use more widespread labels whenever appropriate. Thus, the 'suffixal conjugation' of our Arabic varieties was labelled 'perfective', and the 'free state' of Berber was labelled 'absolute'.
>
> Aprioristic *vs* nonaprioristic categorization of morphemes: however, this unification may have undesired side-effects, in that it may erase the language-specific function of those forms. For instance, the use of the label 'marked nominative' for the 'annexed state' of Berber might at first sight be desirable, because it is currently widespread among typologists. But the function of the annexed state of Berber is far more complex than the definition of the marked nominative implies. Therefore, the traditional label was retained, and the reference to case avoided.

The **rx** line was originally a part-of-speech line. But when we started thinking about the queries that such an online corpus was supposed to allow, we realized that parts of speech were only just a small part of the necessary information. We therefore started with the queries themselves, and implemented the **rx** line with all relevant information, regardless of their linguistic domain. We thus also provide complementary morphological information (neutralization or syncretism, morphological verb-class, etc.), as well as syntactic (word-order, etc.) and semantic (stative verb, etc.) information. We are currently testing the **rx** line for all those types of information. If the information load were too high, we might create an additional tier.

The labels used in **rx** are sometimes the same as those used in **ge**. But they cover a different domain. For instance PREP in **ge** is a special prepositional

paradigm of pronominal affixes, that is found in Berber, Semitic and Chadic. For instance, the third singular pronominal affix attached to the preposition 'on' will be glossed "on-PREP3.SG", and translated "on him". The prepositions in **ge** are glossed by their value only (either grammatically, e.g. LOC, or semantically, e.g. between). In **rx**, PREP means that the morpheme is a preposition. This is useful for specific queries, because sometimes, the same morpheme can be a preposition, or a conjunction.

Here is an example of query: "search ANN in ge & cov in rx" will give us a concordance listing all the examples (with context) where a noun which does not morphologically mark the distinction between the two states (cov= no overt distinction) is (covertly[12]) in the annexed state (ANN). The usefulness of the query lies in the fact that the distinction in Kabyle is covert for half the nouns in texts, therefore it may be interesting to retrieve all those cases, and see what their statistical distribution is: as postverbal subject, nominal modifier, complement of prepositions, etc.

Finally, the **ft** line was apparently unproblematic, but eventually raised some questions since it appeared that indexation to minor units was only possible in some languages, while others were better translated within broader units (major ones). It also appeared that translating a text was in no way an easy task, since contrary to the translation of isolated examples for grammatical purposes, text translations must also provide equivalences for pragmatic dimensions.

We are also planning on adding another tier synchronized to time which will fuse minor units into major units. This would allow to listen to a major unit instead of only minor units, when, for example, the latter is too short to understand the meaning of the sequence. Another free translation tier corresponding to those longer units will be added too.

3. The metadata

In relation to the previous point, translations often contain a certain amount of implicit information, which might be difficult to retrieve for a linguist who did not participate in the recording. This type of information, as well as other types, should be contained in the metadata accompanying the corpus.

There are two types of metadata: one is linked to the technical characteristics and status of the audio recording, the other to the texts themselves. The latter must at the same time provide all the necessary information for the texts to be anchored and understandable to an outsider who was not present during the

[12] 'Covertly' meaning here that if another noun, which has two forms (one for the absolute and one for the annexed) was in the same context, this noun would be in its annexed form. Therefore a noun that does not mark the distinction because it is a borrowing, will be glossed as ANN (or ABS) in ge. And the fact that the word does not mark the distinction will be glossed as "cov" in rx.

recording, and protect the recorded speakers from any prejudice. In that view, as the data is to be made available online to the community, a thorough reflection process was engaged before data collection, concerning the deontological aspects of the project. Thus, anonymization procedures, as well as control over sensitive data (restricted access), have been implemented. In this process, we followed international recommendations, stated in *Corpus Oraux, Guide des Bonnes Pratiques*[13] (Baude 2006).

At the same time, all the relevant information was listed, in order to provide rich metadata on the recordings. These metadata follow the requirements of OLAC[14] (Open Language Archives Community). We provide in annex the metadata form we have devised for each recording.

Conclusion

Two years after the beginning of the CorpAfroAs project, we are able to present a layout (the "CorpAfroAs format"), with a series of organized tiers, and a number of transcription and glossing rules, as well as a list of glosses for the *ge* and *rx* tiers, and a metadata form. Minor alterations will be made in the next two years, but the format is bound to remain quite similar to what it is now. The remaining work consists in finishing the annotation of the data, and working on the queries, theoretically as well as technically. The development of a powerful interface to query a selected sub-corpus and choose different end-user visualizations of the results is also part of the remaining tasks.

References

Baude O. (ed), 2006, *Corpus Oraux, Guide des bonnes pratiques*, CNRS, Paris.

Mettouchi A., Caubet D., Vanhove M., Tosco M., Comrie B. & Izre'el S., 2010, CORPAFROAS, A Corpus for Spoken Afroasiatic Languages: Morphosyntactic and Prosodic analysis, in *CAMSEMUD 2007. Proceedings of the 13th Italian meeting of Afro-Asiatic Linguistics (Udine, May 21st-24th, 2007)*, Edited by Frederick Mario Fales & Giulia Francesca Grassi, Padova, Editrice e Libreria S.A.R.G.O.N.

Wichmann A. 2000, *Intonation in Text & Discourse: Beginnings, Middles and Ends*, Longman, Harlow.

[13] http://www.culture.gouv.fr/culture/dglf/Guide_Corpus_Oraux_2005.pdf

[14] http://www.language-archives.org

CorpAfroAs Metadata

Material description of the archive

Collector (First_name Family_name)

Date-gathering date (yyyy-mm-dd)

Data-gathering place (country, area, village...)

Data type ◉audio ○video

Record characteristics: Recording device : Microphone : Sampling rate (Hz) : Sampling depth (bits) :

Record duration (hh:mm:ss)

Participants

Speaker1 (First_name Family_name)

Informations: age, sex, dialect, ethnic group, birth place, profession, linguistic competence (bilingual (detail), monolingual), locutor comments

Anonymisation

Speaker2 (First_name Family_name)

Informations: age, sex, dialect, ethnic group, birth place, profession, linguistic competence (bilingual (detail), monolingual), locutor comments

Anonymisation

Relations between locutors: family ties, professional ties, etc.

Linguistic description of the archive

Discourse type ◉narration ○conversation

Specify: Tale, story, interview...

Language (code/name)

Title:

Secondary title

Description(s): summary, etc.

Keywords (word1, word2,...):

Working language en

Management of the archive

Publisher(s) CorpAfroAs

Rights http://creativecommons.org/licenses/by-nc-sa/2.5/

Audio Filename (.wav):

ELAN Filename (.eaf):

Reset

Control

Send to Christian