מקומה של העברית המדוברת במאגרי לשון השוואתיים

הכינוס הבינלאומי של האקדמיה ללשון העברית העברית החדשה : תמונת מצב
כינוס מקוון 3-1 במרץ 2022

אילאיל יציב מליבר
המכון לחקר שפות המזרח ותרבויותיו, אינלקו, פריז, צרפת

בעשור האחרון הוקמו שני מאגרים אירופיים שבהם העברית המודרנית המדוברת משמשת לצד שפות מדוברות אחרות במאגרי לשון רב לשוניים.

מאגרים אלה זמינים ונגישים במרשתת לכל קהילת חוקרי הלשון.

- CorpAfroas Corpus of AfroAsiatic languages
- CorTypo = Designing Spoken Corpora for Cross-Linguistic Research

המאגרים הללו הוקמו במיוחד לצורכי השוואה בתוך משפחת שפות ומחוץ למשפחה

מאגר לשון השוואתי

- מאגר השוואתי אינו מסד נתונים או ארכיון להבדיל, למשל, מהארכיון של אוניברסיטת היידלברג: (http://www.semarch.uni-hd.de)
- העבודה על המאגר מעלה שאלות מתודולוגיות ותיאורטיות על עצם תכנון המאגר
- מאפשר לערוך השוואה טיפולוגית בין שפות השייכות לאותה משפחה או שונות
- מחייב איסוף טקסטים מדוברים בעלי אופי משותף

במקרה של קורפאפרואס : שיחות ונרטיבים של שעה

- העבודה על המאגר מאפשרת לשאול שאלות על שימוש בתגיות זהות לתיוג חלקי הדיבר, על אופן החיתוך ליחידות פרוזודיות
- מאגר חי ומתפתח
- מבקש לשמש מודל למאגרים השוואתיים נוספים

מאגר לשון השוואתי של שפות אפרו-אסיתיות

(A Spoken Corpus for AfroAsiatic Languages

Prosodic Segmentation and Morphosyntactic Analysis)

Get ADOBE FLASH PLAYER

CORPAFROAS, A CORPUS FOR SPOKEN AFROASIATIC LANGUAGES: PROSODIC AND MORPHOSYNTACTIC ANALYSIS

Objectives of the project

CorpAfroAs is an integrated pilot project realized by field linguists for field linguists and typologists, which proposes:

- A methodology for the treatment of fieldwork textual data in underdescribed languages, from data gathering to automatic searches on the corpus,
- A free, open-source and user-friendly new software, ELAN-CorpA, developed within our project from Elan (Max Planck Institute Nijmegen),
- A pilot corpus composed of annotated first-hand transcriptions of narrative and conversational data in twelve AfroAsiatic languages (one hour per language), with accompanying sound files, list of glosses, grammatical sketches, and metadata..

Objectifs du projet

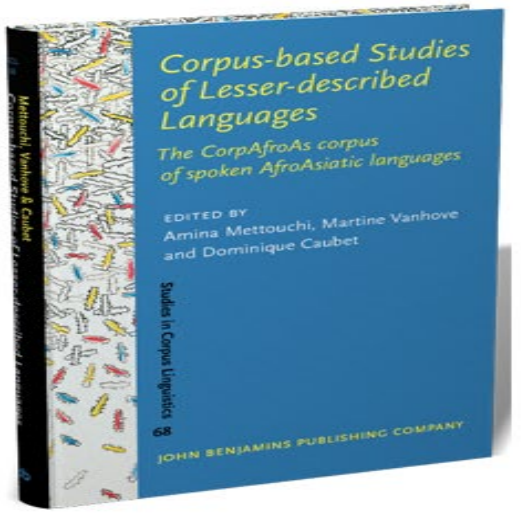
CorpAfroAs est un projet financé par l'Agence Nationale de la Recherche (France), pour 2007-2012. C'est une entreprise unique, en ce qu'elle a permis de mettre à disposition le premier corpus de langues afro-asiatiques (chamito-sémitiques) comportant une indexation texte-son, et une annotation complexe.
Le corpus est librement accessible, et est accompagné par un logiciel, des outils et des publications visant à faciliter la contribution d'autres linguistes de terrain à CORPAFROAS, ainsi que la mise en place d'initiatives inspirées de ce modèle.

DOWNLOAD

ELAN-CorpA

v. 5.7

PUBLICATIONS



הפִּילוּם האפרו-אסיאתי (חמי-שמי): משפחות הלשונות ה**שמיות**, ה**בֶּרבֶּריות**, ה**כושיתיות**, ה**אומותיות** וה**צ'אדיות**; מצרית; אונגותה

מאפייני קורפאפרואס

- מאגר לשון השוואתי חופשי להורדה, למחקר ולהעשרה.

הקורפוס מונגש על שרת, ותוכנה מותאמת מאפשרת חיפושים המצליבים פרטי מידע

- 13 שפות השייכות לפילום האפרו-אסייתי

השפות השמיות : ערבית מרוקאית, ערבית לובית, עברית

השפות הברבריות: תקבילית, תאמאשק, סיווי

השפות הצ׳אדיות : האוסה, זאר

ערבית קריאולית: ג׳ובה

השפות הכושיתיות: ב׳זה, גאוואדה, צאמאקו

השפות האומוטיקיות: וולאיתה

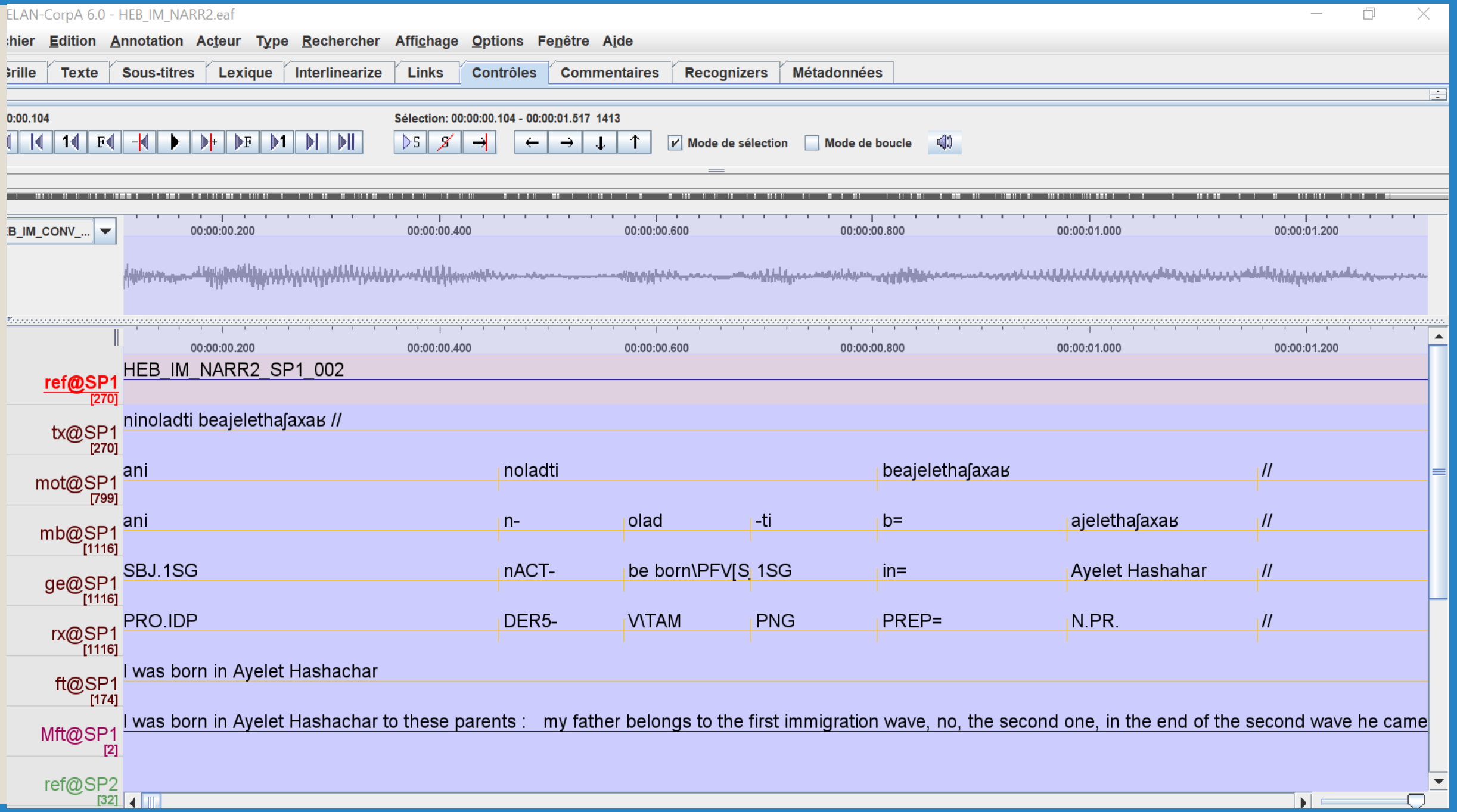
תת מאגר העברית המדוברת

שעה של שיחות ונרטיבים

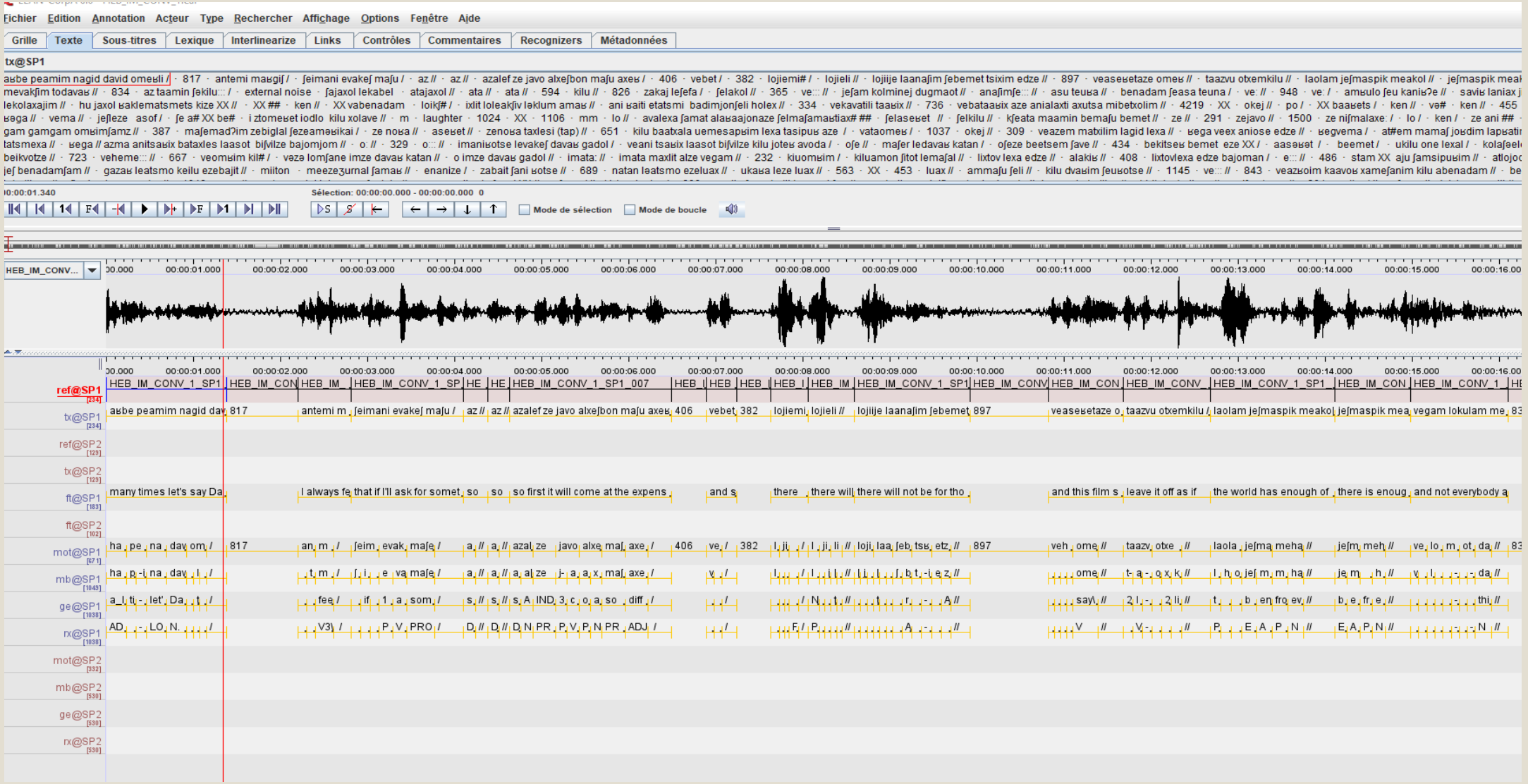
שימוש בגלוסות לתיוג המורפמות המשמשות במחקר על השפות השמיות

@Glossing in Semitic languages: A comparison of Moroccan Arabic and Modern Hebrew (Á. Vicente, I. Yatziv-Malibert and A. Barontini)

שפת יום-יום לא רשמית, חצי ספונטנית


ELAN-CorpA 6.0 - HEB_IM_NARR2.eaf
Chier Edition Annotation Acteur Type Rechercher Affichage Options Fenêtre Aide
Grille Texte Sous-titres Lexique Interlinearize Links Contrôles Commentaires Recognizers Métadonnées
0:00.104
Sélection: 00:00:00.104 - 00:00:01.517 1413
Mode de sélection
Mode de boucle
EB_IM_CONV_...
00:00:00.200
00:00:00.400
00:00:00.600
00:00:00.800
00:00:01.000
00:00:01.200
ref@SP1 [270]
HEB_IM_NARR2_SP1_002
tx@SP1 [270]
ninoladti beajelethaʃaxaʁ //
mot@SP1 [799]
ani noladti beajelethaʃaxaʁ //
mb@SP1 [1116]
ani n- olad -ti b= ajelethaʃaxaʁ //
ge@SP1 [1116]
SBJ.1SG nACT- be born\PFV[S 1SG in= Ayelet Hashahar //
rx@SP1 [1116]
PRO.IDP DER5- V\TAM PNG PREP= N.PR. //
ft@SP1 [174]
I was born in Ayelet Hashachar
Mft@SP1 [2]
I was born in Ayelet Hashachar to these parents : my father belongs to the first immigration wave, no, the second one, in the end of the second wave he came
ref@SP2 [32]

ייצוג הטקסטים: תוכנת ELAN



tx: text תעתיק פוניטי רחב

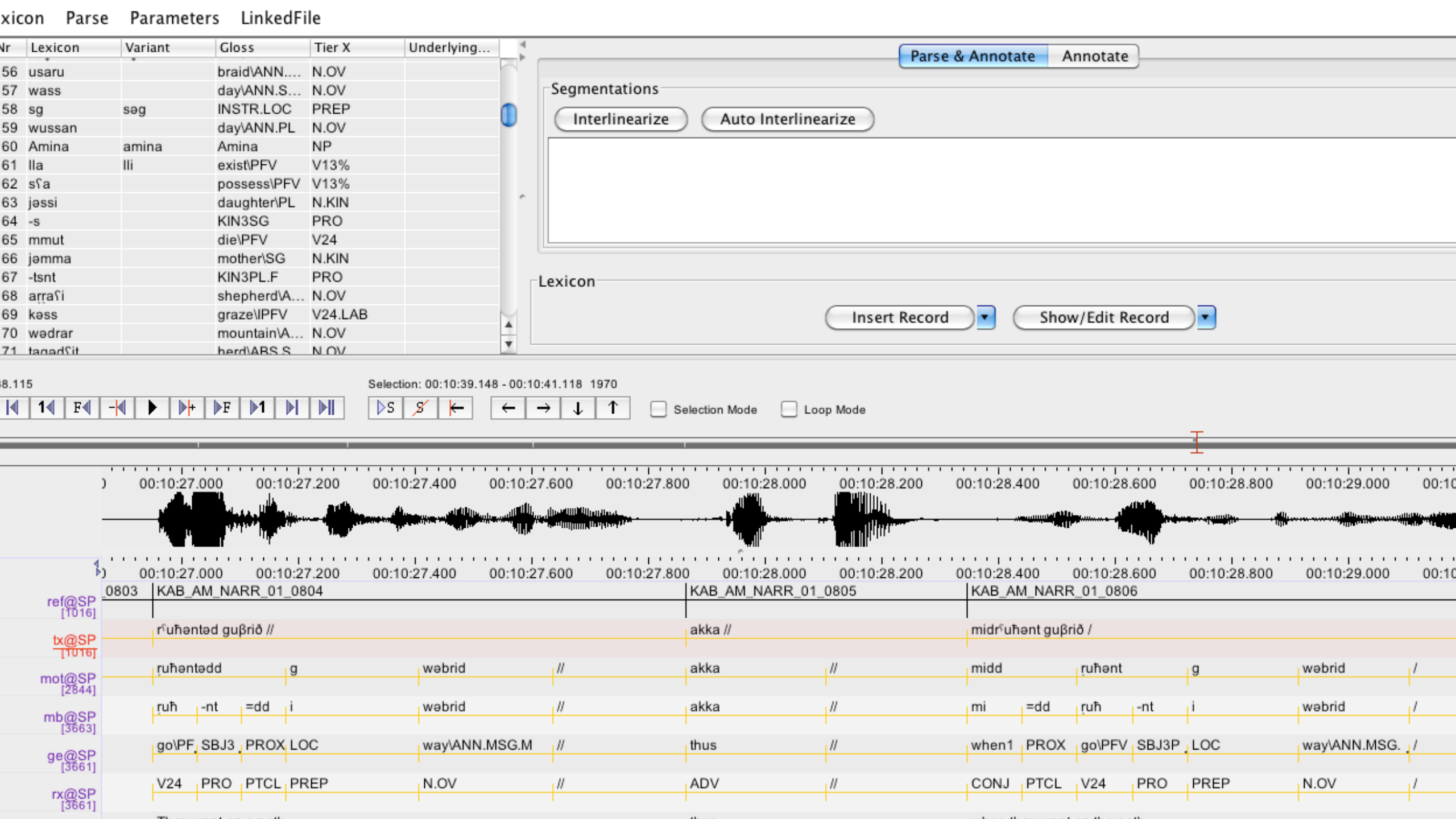
mot: word תעתיק פונולוגי

mb: morphological boundary ניתוח מורפולוגי

ge: glose גלוסות

rx: חלקי דיבר וסימונים שונים

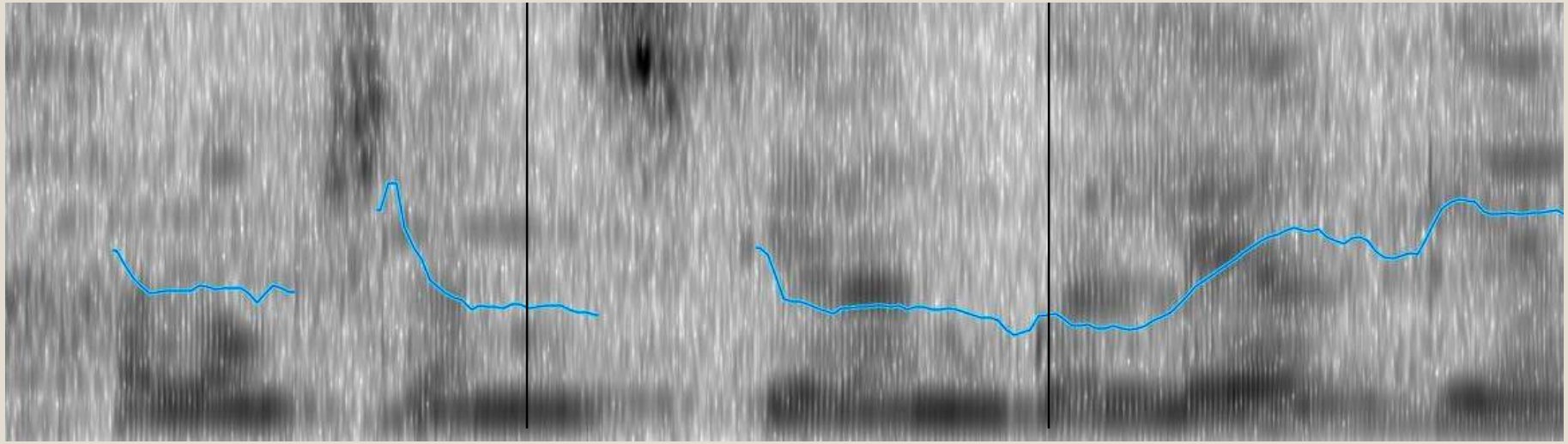
ft: free translation תרגום לאנגלית


xicon Parse Parameters LinkedFile
Nr Lexicon Variant Gloss Tier X Underlying...
56 usaru braid\ANN.... N.OV
57 wass day\ANN.S... N.OV
58 sg səg INSTR.LOC PREP
59 wussan day\ANN.PL N.OV
60 Amina amina Amina NP
61 lla lli exist\PFV V13%
62 sʕa possess\PFV V13%
63 jəssi daughter\PL N.KIN
64 -s KIN3SG PRO
65 mmut die\PFV V24
66 jəmma mother\SG N.KIN
67 -tsnt KIN3PL.F PRO
68 arraʕi shepherd\A... N.OV
69 kəss graze\IPFV V24.LAB
70 wədrar mountain\A... N.OV
71 tagadʕit herd\ABS.S N.OV
Parse & Annotate
Annotate
Segmentations
Interlinearize
Auto Interlinearize
Lexicon
Insert Record
Show/Edit Record
Selection: 00:10:39.148 - 00:10:41.118 1970
Selection Mode
Loop Mode
00:10:27.000 00:10:27.200 00:10:27.400 00:10:27.600 00:10:27.800 00:10:28.000 00:10:28.200 00:10:28.400 00:10:28.600 00:10:28.800 00:10:29.000
0803 KAB_AM_NARR_01_0804 KAB_AM_NARR_01_0805 KAB_AM_NARR_01_0806
ref@SP [1016]
tx@SP [1016]
mot@SP [2844]
mb@SP [3663]
ge@SP [3661]
rx@SP [3661]
rʕuħəntəd guβrið // akka // midrʕuħənt guβrið /
ruħəntədd g wəbrid // akka // midd ruħənt g wəbrid /
ruħ -nt =dd i wəbrid // akka // mi =dd ruħ -nt i wəbrid /
go\PF SBJ3 PROX LOC way\ANN.MSG.M // thus // when1 PROX go\PFV SBJ3P LOC way\ANN.MSG. /
V24 PRO PTCL PREP N.OV // ADV // CONJ PTCL V24 PRO PREP N.OV /

סגמנטציה פרוזודית: יחידות פרוזודיות וייצוגן

שורת tx ושורת mot: מילים פרוזודיות ומילים מורפוסינטקטיות

עברית:



tx: χalomʃlanu zʃtijelanu galeʁgja

mot: χalom ʃelanu ze ʃe tihje lanu galeʁgja

mb: χalom ʃel=anu ze ʃe=t-ihje l=anu galeʁgj-a

ge: dream of=POSS.1PL DEM.SG.M NMNL=3SG.F-be\NFCTto=POSS.1PL gallery-F

ft: "Our dream is that we will have our (own) gallery."

תוצרי המאגר

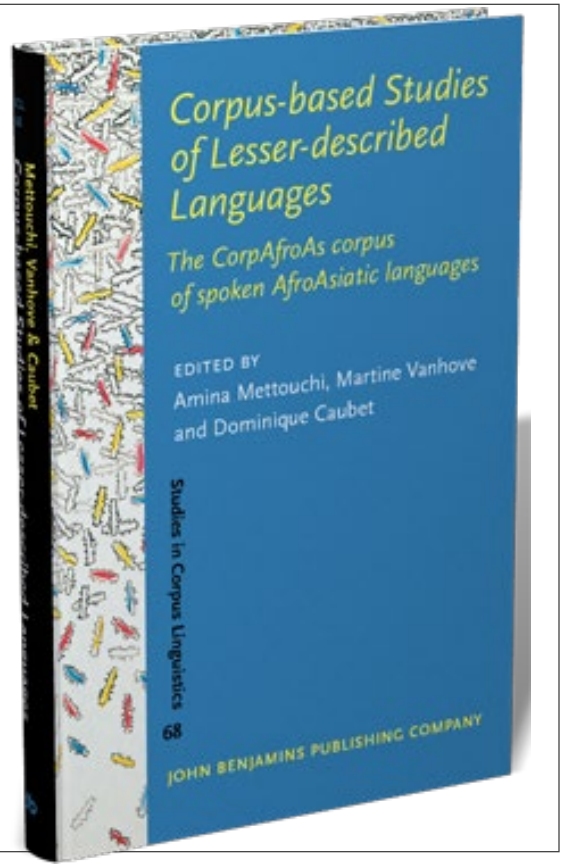
- מדריך מפורט למשתמש העתידי החל מהריאיון לצורכי הקלטה וכלה בתיוג באמצעות תוכנת Elan-CorpA
- פיתוח של תוכנת Elan לצורכי תיוג וחיתוך ליחידות פרוזודיות וגם לצורכי חיפוש מתקדמים
- אתר המאפשר שימוש חינמי בכל ההקלטות
- פרסום קובץ מאמרים של משתתפי הפרויקט
- פרויקט טיפולוגי נוסף שבבסיסו המאגר החלוצי של קורפאפרואס

פרסומים

Corpus –bases studies of lesser-described languages

The corpAfroAs corpus of spoken afroasiatic languages

Edited by A.Mettouchi, M. Vanhove & D Caubet



תגליות שעלו מן ההשוואה

Quotative constructions and prosody in some Afroasiatic languages: Towards a typology (Malibert & Vanhove)

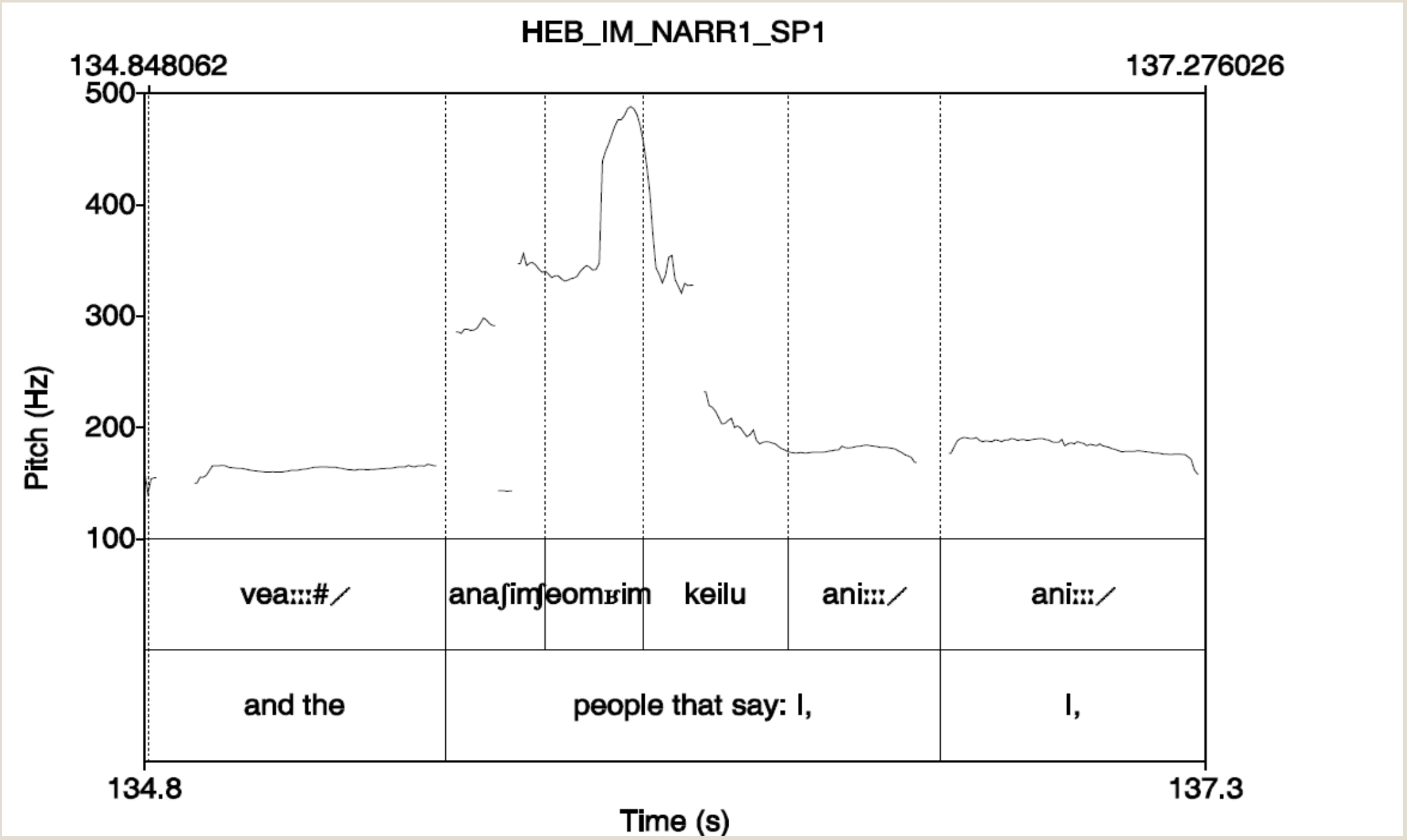
ניתוח טיפולוגי של הממשק של דיבור ישיר ועקיף ופרוזודיה מתבסס על המונח עקומת ההטמעה הפרוזודית.

ההטמעה הפרוזודית מאופיינת על ידי:

- ❖ מקומם של הגבולות הפרוזודיים ביחס לפועל הדיבור
- ❖ דגם העקומה שבסופה גבול סופי
- ❖ שינויים בפיץ', בעוצמת הקול, מנעד ואורך

העקומה נעה על הרצף שבין **הטמעה מושלמת** של ''פסוקית הדיבור'' בתוך מבנה הציטוט לבין **עצמאות פרוזודית מוחלטת**

(35) *ve = a:::# /*
and=FS
*anaʃim ʃe = omʁ-im keilu **ani::: /***
men\PL COMP=say\PTCP.ACT-PL.M like SBJ.1SG
ani::: /
SBJ.1SG
lemaʃal a# /
for_example FS
lemaʃal kʃe = ata mekabel stam //
for_example when=SBJ.2SG.M obtain\PTCP.ACT[SG.M] whatever
notn-im dugm-a /
give\PTCP.ACT-PL.M example-SG.F
'and the people who say like: I, I, for instance I… for instance when you get… they give there an example… '(HEB_IM_NARR1_SP1_138-143)


HEB_IM_NARR1_SP1
134.848062
137.276026
500
400
300
200
100
Pitch (Hz)
vea:::#/
anaʃimʃeomʁim
keilu
ani:::/
ani:::/
and the
people that say: I,
I,
134.8
137.3
Time (s)

המאגר ההשוואתי השני, המשכו של קורפאפרואס, ממשיך לשאול שאלות על עיצוב מאגר לשונות דבורות לצורכי מחקר השוואתי של **שפות מרוחקות זו מזו**

Designing spoken corpora for cross-linguistic research

ההשוואה מתאפשרת הודות לקטגוריות אמפיריות הנובעות מניתוח ותיוג כל שפה לחוד, וצומחת מלמטה כלפי מעלה (BOTTOM-UP)

Kabyle
Hebrew
Amis
Ixcatec
Wolof
Bedja
Wandala
Wolaitta
Zaar
Gbaya
Pévé
Movima

CorTypo:
Designing spoken corpora for cross-linguistic research

CorTypo is a pilot research project (linguistic typology) funded by the French Agence Nationale de la Recherche and aimed at testing hypotheses about the similarities and differences between the languages of the world.

Coordinated by Amina Mettouchi at UMR 8135 of the CNRS (LLACAN), its team is composed of four computer scientists and twelve international researchers including Zygmunt Frajzyngier, author of the theoretical approach implemented.

The project started in March 2013 and lasted 48 months. It received €230,000 of ANR funding, for an overall cost of about €2,730,400.

תודה רבה