Mettouchi, Amina, Martine Vanhove and Dominique Caubet (2015) "Preface". In: Mettouchi, Amina, Martine Vanhove and Dominique Caubet (eds.), Corpus-based Studies of Lesser-described Languages: The CorpAfroAs corpus of spoken AfroAsiatic languages. Studies in Corpus Linguistics 68. John Benjamins: Amsterdam-Philadelphia. vi, 332 pp. + index (pp. 1–9)

Preface

Amina Mettouchi, Martine Vanhove and Dominique Caubet
EPHE (LLACAN) / CNRS (LLACAN) / INALCO (LaCNAD)

When the CorpAfroAs project was submitted to the French *Agence Nationale de la Recherche* in 2006, there was only one website, *The Semitisches Tonarchiv* providing online data in Afroasiatic laguages, in the form of sound files accompanied by transcriptions in pdf format, and another project, the Corpus of Spoken Israeli Hebrew (CoSIH), which was at a standstill, with no available data online. Other language families were more largely represented on the web, in the form of Archives at LACITO, DoBeS or ELDP, among others.

At the time, however, even the richest of repositories in lesser-described languages had not integrated systematic prosodic segmentation in their transcription of the data. And none had chosen a systematic annotation schema in view of typological research.

In this context, CorpAfroAs appeared as a pioneering endeavor, a status that is still valid at the time we are releasing the data and publishing the accompanying volume.

The project involves the collection of one hour of spontaneous speech per language, 60% monologal and 40% dialogal, in thirteen Afro-Asiatic languages; the sound-indexed transcription, annotation, and translation into English of those thirteen subcorpora; the elaboration of grammatical sketches for each language; and the development of a lexicon-assisted annotation tool in the software ELAN named ELAN-CorpA. The aim of the project is not only to provide data, but to offer a methodology for the creation of corpora in lesser-described languages, from data collection, through analysis, to online dissemination. All stages of the process have been documented in a Manual available online at http://dx.doi.org/10.1075/scl.68.website.

CorpAfroAs is characterized by its integrated dimension:

- a common layout for the annotation of sound files,
- a unified list of abbreviations, allowing searches across languages,
- an accompanying grammatical sketch per language, where glosses are given language-internal definitions.

The corpus is searchable online, within and across languages. Ultimately, the pilot corpus is designed to grow and become a reference corpus, as well as to inspire initiatives for other language phyla.

The languages represented in the online corpus are:

Kabyle, Tamashek (Berber),
Hausa and Zaar (Chadic),
Afar, Beja, Gawwada, Ts'amakko (Cushitic),
Wolaytta (Omotic),
Moroccan and Libyan Arabic, Hebrew (Semitic),
Juba-Arabic, an expanded Arabic-based pidgin.

In its pilot form, the corpus is not designed to present a balanced sample of languages. It covers all branches, and different types of languages, in order to provide technical and scientific solutions for all potential types: tonal and intonational, concatenative and non-concatenative, endangered as well as rather well-described languages, with or without codeswitching, etc.

The project is organized along two axes, prosody and morphosyntax, which are linked to the nature of the materials and to the aim of the project, namely crosslinguistic comparability.

Our first research question bears on the prosodic structure of the languages of the project, and more precisely, on the type of segmentation relevant for our data. This task is one of the main innovative aspects of our project. We decided to index the recordings on intonation units, a level that is largely recognized as useful for grammar, discourse and conversation analysis, and which was for instance used in the C-ORAL-Rom corpus of spoken Romance languages, developed by Cresti and Moneglia (2005).

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

The second research question concerned the morphosyntactic organization of the languages of the corpus. The corpus is not only translated, but also interlinearly glossed. For this purpose, we have developed a format allowing several annotation tiers. The purpose of those tiers is the automatic retrieval of a number of relevant queries concerning Afroasiatic languages: pronominal systems, case systems, nominal predicates, aspect, ideophones, demonstratives, verbal derivation, etc. Here are the various tiers and their contents:

ref identifier for the annotation unit (time-aligned)
tx transcription in broad phonetics into phonological words (Symbolic Association)
mot intermediary tier allowing the segmentation into morphosyntactic words (Symbolic Subdivision)
mb morphophonological transcription into morphemes (Symbolic Subdivision)
ge morpheme-by-morpheme gloss of mb according to the Leipzig Glossing Rules, expanded within the project (Symbolic Association)
rx part-of-speech and other information relevant for retrieval purposes (Symbolic Association)
ft or mft free translation into English (Symbolic Association)

The annotation system we chose is based on the Leipzig Glossing Rules, developed jointly by the Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology (Bernard Comrie, Martin Haspelmath) and by the Institute of Linguistics of the University of Leipzig (Balthasar Bickel), in order to promote convergence in glossing systems. The list of morphemes being open, and the rules devised more for readability than for automatic retrieval, one of the tasks we achieved is the establishment of a completed list, with rules adapted to computer retrieval. The full list is available on line on the project website at http://dx.doi.org/10.1075/scl.68.website, and it is possible to suggest additions through the use of an online form.

Each member of the project collected, transcribed, segmented and annotated the data in their language, on the basis of language-internal consistency, but also in view of cross-linguistic comparison. Our main concern was to find the optimal degree of unification of the annotations, in order to both respect the specificities of languages, and provide a comparative basis for typology. It turned out in the course of the project that cross-linguistic comparison could not rely directly on the corpora, even with a list of abbreviations and definitions, but had to be mediated through grammatical sketches, which were therefore added to the deliverables of the project. Those sketches are online, and give the end-user sufficient insight into the definition of the categories involved in order to find the relevant typological matches in the other languages of the corpus.

Ethical and technical aspects have also been largely dealt with in this project. As the data are made available online to the community, a thorough reflection process was initiated before data collection, concerning the ethical aspects of the project. Thus, anonymization procedures, as well as control over sensitive data, have been implemented when needed. At the same time, all the relevant information was gathered, in order to provide rich metadata on the recordings, in the

IMDI format. These metadata follow the requirements of OLAC (Open Language Archives Community) and the TEI (Text Encoding Initiative). The recordings were all digital, with strict requirements as to the format: non-compressed, .wav files, recorded at 44.1 khz / 16 bits, with high-quality microphones and pre-amplifiers. The high quality of the recording is necessary, not only because one of our scientific aims is to conduct a prosodic analysis of the data, but also for conservation purposes.

The software used for the analysis of the data were Praat for segmentation, and ELAN-CorpA, or Toolbox followed by ELAN for primary annotation. Indeed, before the development of ELAN-CorpA within the project, it was necessary to import the Praat Textgrid into ELAN, then export the resulting file into Toolbox, then annotate in Toolbox, then reimport the Toolbox file into ELAN in order to have a sound-indexed, searchable file. ELAN-CorpA, a lexicon-aided annotation made it possible to do without the complex process of data treatment and annotation via Toolbox. Once the corpus was fully annotated, a website for online queries was devised, in collaboration with another ANR project of our research department, *Sénélangues*. This tool, ELAN-WebSearch, is similar to the search module of ELAN, but is based on a PostgreSQL database, and includes additional functionalities (concordances and lists).

The technical dimension of the project implies of course the participation of an engineer on a permanent basis (Christian Chanard, of the LLACAN research unit), as well as a developer specially hired for the project, Coralie Villes. Regular collaboration with Han Sloetjes of the Max Planck Institute at Nijmegen has guaranteed the adaptation of ELAN-CorpA to the general architecture of ELAN.

Once the corpus was released in its Beta version, it appeared desirable to publish an accompanying volume where various central aspects of the CorpAfroAs project would be developed, and some scientific results published.

The first part of this volume is dedicated to phonetic, phonological and prosodic aspects of the project. Izre'el and Mettouchi's survey of the phonetic and transcriptional aspects of CorpAfroAs sketches a portrait of the Corpus according to the choices that were implemented. First of all, the priority was given to the close relationship between the **tx** tier and the sound file, mirroring the structure of the software, in which **tx** is indexed to the sound file represented by the waveform window in ELAN. The transcription in **tx** was therefore meant to reproduce as faithfully as possible the spoken monologue or interaction, allowing the end-user to recognize the elements of the speech continuum. However, the length of the corpus does not allow detailed phonetic representation, therefore, a degree of phonologization of the transcription was introduced, resulting in a broad phonetic transcription. In this tier, words are phonological (as opposed to morphosyntactic). The segmental string was segmented into prosodic units, defined by their boundaries and by their coherent internal contour. Intonation units were chosen over syntactic units (clauses or phrases) because they are the organic units of spontaneous speech. At a later stage, the corpus can be further segmented into other units if needed for further research on the correspondence between syntactic and prosodic units.

The **tx** tier was in turn further morpho-phonologized so that the **mot** tier should be composed of morphosyntactic words, morphemically transcribed. This level opens the way for a tokenization into morphemes in the **mb** tier. Those morphemes are then glossed in **ge** and **rx**. Finally, a free translation was given, which is currently aligned with respect to Intonation Units in most languages of the corpus, and to larger units (paratones) for the SOV languages of the sample, thus allowing better alignment between the source language and the target language (English) for the translation.

The comparison between **tx** and **mot** allows the systematic study of sandhi and other similar phenomena, and of the syntax/prosody interface. The segmentation into prosodic units allows the study of various interfaces: syntax, information structure, discourse. The paper provides a detailed discussion of the various units of speech, and arguments for the decisions made by the team regarding segmentation and transcription.

Bernard Caron's paper specifically broaches the question of intonation in tone languages, with data from Zaar (Chadic). He shows how pitch plays a role in the intonation of a three-tone language, through the observation of the variations between post-lexical tones as they are perceived and transcribed by native speakers, and their acoustic realisation as represented by Praat and Prosogramme, and how surface tones accounted for and/or predicted by postlexical tonological rules undergo further variations. Those variations fall under two main categories: (a) declination; and (b) intonemes, defined as the minimal units of distinctive intonation contours associated with particular functions (pragmatic, information structure…). These are further divided into terminal intonemes (fall, rise, level and high-rise) and initial intonemes: step-down and step-up. The study allows the author to classify Zaar as a mixed language as regards Bearth's (1998) typology: a language which both stacks intonation patterns over lexico-grammatical tones and expresses intonation at the periphery of the utterance, i.e. with both internal and peripheral intonation.

The second part is dedicated to studies of interfaces between prosody, information structure or syntax. Bernard Caron, Cécile Lux, Stefano Manfredi and Christophe Pereira analyze the correlation between prosody, sentence types, morphology and information structure categories, with a special emphasis on topic, focus and frames (i.e. left-dislocated circumstantials). After having examined into details the various prosodic patterns for four languages of the CorpAfroAs corpus

(Zaar, Tamasheq, Juba Arabic and Tripoli Arabic), they conclude that (i) despite their different phonological pitch systems, and some differences in the correlations between prosodic contours and information structures, strong common tendencies emerge concerning the default intonation patterns of thetic sentences, and of topics (with the exception of Tamasheq); (ii) the variations in the intonation patterns of polar questions, focus and Wh-Questions follow a rule: a lack of a specific intonation pattern for a specific information structure is supplemented by morpho-syntactic marking, in other words the more a structure relies on morpho-syntax, the less it relies on intonation.

Il-Il Malibert and Martine Vanhove's paper investigates, in a crosslinguistic perspective, the relationship between prosodic contours and direct and indirect reported speech (i.e. without or with deictic shift) in four typologically and genetically different Afroasiatic languages of the CorpAfroAs pilot corpus: Beja (Cushitic), Zaar (Chadic), Juba Arabic (Arabic based pidgin) and Modern Hebrew (Semitic). The descriptive tools and analysis of Genetti (2011) for direct speech report in Dolakha Newar (Tibeto-Burman) are used as a starting point and adapted to the annotation system of CorpAfroAs. Each language section investigates the prosodic cues and contours of direct speech reports, in relation to their quotative frame and their right and left contexts. The same prosodic features are also investigated for the three languages in our corpus which have indirect reported speech (Zaar, Juba Arabic and Hebrew). It is shown that speech reporting as a rhetorical strategy varies a lot from one language to another and is more frequent in the three unscripted languages of the sample. Even if speech reports show a wide range of prosodic behaviors, there are nonetheless clear tendencies that become apparent and which are related to various factors: speech report types, types of constituents of the quotative frame, genres, and typological features of the languages in question. A preliminary typology of the interface between prosody and speech reporting is proposed.

The third part of the volume deals with the problem of cross-linguistic comparability. This issue is relevant on several levels: the choice and type of glosses, the role of the accompanying grammatical sketches, the kinds of queries allowed by the chosen layout, which is based not on one, but on two annotation tiers: **ge**, and **rx**. The **ge** line provides morpheme-by-morpheme glossing in terms of the function of each form. The **rx** line is devoted to part-of-speech information, as well as any gloss that the researcher considers helpful for the retrieval of relevant phenomena in the corpora (syncretism, verb class, noun class, syntactic relations, etc.).

Angeles Vicente, Il-Il Malibert and Alexandrine Barontini show how a project such as CorpAfroAs can challenge existing traditions in terms of annotation, and develop proposals for a standard in the domain of Semitic linguistics. Indeed,

Semitic studies traditionally used no interlinear glossing, translated examples were the norm, with occasional morphological information (SG, PL...) interspersed inside word-by-word translations. Those habits hamper the diffusion of Semitic studies outside the circle of language specialists. They also prevent the development of morphosyntactically-annotated corpora, and possible family-internal comparisons. The authors of the paper provide a series of annotation proposals that take into account the complex morphology characterizing this language-family. This endeavor also paves the way for the comparison of similar categories in Moroccan Arabic and Spoken Hebrew.

The paper by Comrie addresses the issue of glossing examples in unfamiliar languages, with particular reference to the CorpAfroAs project. He first introduces the Leipzig Glossing Rules (LGR), which were devised with a very specific purpose in mind, namely to standardize the notations used by linguists, especially typologists, in presenting the morphological structure of example sentences in languages unfamiliar to the reader. While LGR form a suitable basis for annotation in projects like CorpAfroAs, such projects have a higher level of requirements, in particular the need to be able to retrieve particular categories and structures from corpora in various languages. The article discusses with examples the enrichment of LGR that is needed for this purpose, in particular the addition of extra tiers to capture such notions as part of speech and grammatical relation.

Amina Mettouchi, Graziano Savà and Mauro Tosco's paper test the potential for cross-linguistic comparability of CorpAfroAs through the study of three morphosyntactic phenomena represented in several languages of the corpus ('ventive' extensions, gender, and case); they show that CorpAfroAs indeed allows the retrieval of a body of data amenable to cross-linguistic comparison, within the AfroAsiatic phylum and beyond, but that, given the annotation scheme of the corpus, the retrieval of relevant data also relies on information given in the accompanying grammatical sketches. This being taken into account, the paper proves that such automatic cross-linguistic investigations are powerful enough to allow the retrieval of complex data for gender; they can provide quite precise characterizations of lexical information concerning the types of verb that co-occur with directional extensions; and they contribute to the debate on the function of case-marking and syntactic roles in several languages of the corpus. In this respect, CorpAfroAs, as a pilot corpus, can indeed serve as a basis for typological investigations.

Zygmunt Frajzyngier and Amina Mettouchi's paper carries further the discussion on cross-linguistic comparability, by showing how the corpus can be completed by a comparative database that allows empirical analysis of the data. In this proposal, comparison is no longer based directly on the labels used for annotation, but it is mediated by the establishment, by the author of the corpus and specialist of the language, of the functional domains of that language. And it is those domains that are compared, along several dimensions: type and number of oppositions, formal means used for the encoding of functions and predications. The domain considered for illustration is that of Reference, in Kabyle (Berber) and Mina (Chadic). This proposal is currently being implemented in a new ANR-funded project coordinated by Amina Mettouchi, *CorTypo*.

The fourth part concerns two different contact-induced phenomena found in several languages of the CorpAfroAs corpus: codeswitching and borrowing. Stefano Manfredi, Marie-Claude Simeone-Senelle and Mauro Tosco provide new linguistic evidence for the identification of codeswitching in contrast to lexical borrowings. The innovative dimension of this paper lies in the use of prosodic information for the analysis of those phenomena. The authors show that variation in intonation contours provides a good test for telling them apart. They show in addition that different types of codeswitching have different prosodic segmentation patterns: intersentential codeswitching is systematically related to monolingual intonation units, while intrasentential codeswitching tends to occur at the end of bilingual intonation units; on the other hand tag-switching is regularly highlighted by prosodic prominence. This result can be taken as a new constraint for distinguishing codeswitching from borrowing.

The volume ends with the presentation, by Christian Chanard, of the software development conducted within the CorpAfroAs project on the basis of the software ELAN (Max Planck Institute, Nijmegen). This development, resulting in the ELAN-CorpA software, involves the addition of an internal parser linked to a lexicon, for semi-automatic interlinearization purposes. This addition was made necessary due to several obstacles encountered in the course of the project realization, namely lack of text-to-sound indexation in Toolbox, the most widely-used software in African field linguistics, problems of compatibility for Mac-users, lack of complex search functions in current software. ELAN-CorpA is a good example of the way bottom-up proposals made by field linguists in lesser-described languages can make their way into information technology, thus facilitating the treatment of data, both at the source, and for the end-user.

In conclusion, the collection of papers in this volume is a quite unique endeavor in that it provides data, analyses and discussions on several aspects of corpus linguistics that are rarely studied together. By combining research on lesser-described languages, cross-linguistic comparison within a phylum, the introduction of prosodic segmentation and several layers of transcription and annotation, the use of first-hand dialogal and monologal recordings, the elaboration of accompanying grammatical sketches and standardized glosses, CorpAfroAs provides a rich integrated basis for corpus-based and corpus-driven investigations.

Acknowledgements

We are most grateful to our colleague David Roberts who reviewed the volume from the point of view of a native English speaker.

References

The Semitisches Tonarchiv. <www.semarch.uni-hd.de/>

Corpus of Spoken Israeli Hebrew (CoSIH). <http://humanities.tau.ac.il/~cosih/english/>

Archives (now named *Pangloss*) at LACITO. <http://lacito.vjf.cnrs.fr/archivage/>

DoBeS. <www.mpi.nl/resources/data/dobes>

ELDP. <www.hrelp.org/archive/>

Mettouchi, Amina, Vanhove, Martine & Caubet, Dominique (eds). *Corpus-based Studies of Lesser-described Languages: The CorpAfroAs Corpus of Spoken AfroAsiatic Languages.* <http://dx.doi.org/10.1075/scl.68.website>

ELAN-CorpA. <http://dx.doi.org/10.1075/scl.68.website>

CorpAfroAs Manual. <http://dx.doi.org/10.1075/scl.68.website>

CorTypo. <http://cortypo.huma-num.fr/>

Bearth, Thomas. 1998. Tonalité, déclinaison tonale et structuration du discours - Un point de vue comparatif. In *Les unités discursives dans l'analyse sémiotique: La segmentation du discours,* Gustavo Quiroz, Ioanna Berthoud-Papandropoulou, Evelyne Thommen & Christina Vogel (eds), 73–87. Bern: Peter Lang.

Cresti Emanuela & Moneglia, Massimo (eds). 2005. *C-ORAL-ROM : Integrated Reference Corpora for Spoken Romance Languages* [Studies in Corpus Linguistics 15]. Amsterdam: John Benjamins. DOI: 10.1075/scl.15