International Journal of Modern Anthropology
Int. J. Mod. Anthrop. (2015) 8: 94 - 116
Available online at: www.ata.org.tn ; doi: http://dx.doi.org/10.4314/ijma.v1i8.5

Original synthesis Article

The evolution of Arabic(s)
Making the Idiom speak for the Deme

Mahé Ben Hamed[1*], Mélissa Barkat-Defradas[2], Rim Hamdi-Sultan[3]

[1]*Laboratoire Bases, Corpus, Langage (UMR 7320 – Nice);* [2]*Institut des Sciences de l'Évolution de Montpellier (UMR 5554 – Montpellier);* [3] *Laboratoire Savoir, Textes et Langage (UMR 8163 - Lille)*
[*] *Corresponding author, e.mail: mahe.ben-hamed@umontpellier.fr*

Abstract - Despite its rather shallow origin, Arabic forms the largest group of extant Semitic languages and one of the most geographically widespread languages of the world. The current distribution of its linguistic variants is the product of a phylogeography of the populations that spoke them, and Arabic dialects have captured in their words and structures traces of their speakers demic history. In this paper, we show how a phylolinguistic approach can identify such traces and make sense of them in terms of population contacts and migration, and discuss how its findings fit with the cumulative knowledge of the history and genetics of arabic-speaking populations.

Keywords **:** Phylo-linguistics, cultural evolution, modern synthesis, Arabic dialects, Afro-Asiatic, phylogenetic networks

The aim of this paper is to explore what the linguistic diversity of Arabic dialects has to say about the history – past and recent – of their speakers. Across the Arabic-speaking geolinguistic domain, dialects exhibit a relatively high diversity given their short span of evolution – some 1400-1600 years, which is partly due to language-internal factors, but also the demic history of speech communities. Migration and contact impact both a population's genetic *and* linguistic makeups, and the challenge lies in making sense of the observable linguistic diversity in terms of population history.

In the 1980's and 1990's, Cavalli-Sforza's claims of gene-language co-evolution (Cavalli-Sforza et al., 1988 ; 1992) sparked growing interest in correlating human population genetic data to the populations linguistic characteristics (see Ben Hamed and Darlu, 2007 for a review). The years 2000 tend to reverse the gene-language comparative paradigm by correlating the evolution of languages to the demic and cultural history underlying it (see Gray et al., 2010 ; Ben Hamed, 2015). Like genes in the biological realm, words fossilize in their structure traces of language evolution, and like genes, words can be analyzed within a computational phylogenetic framework.

This paper will put a phylogenetic analysis of Arabic dialects inter-lexical divergence in perspective, on one hand with Arabic historical dialectology and on the other with the demic history of Arabic populations. After presenting the geolinguistic landscape of Arabic and the state of affairs in historical dialectology, we will move on to the phylogenetic network approach that will allow a detailed exploration of the genetics and admixture of languages, in correlation to the demic history of their speakers. We will conclude by discussing the scope of such a synthetic dialectological approach beyond the case of Arabic, as well as its limits and perspectives.

Arabic – a contrasted geolinguistic landscape

Of all Afroasiatic languages, Arabic is the only one true to its family name (Fig.1). While all non-Semitic but also non-Arabic Semitic languages are located in either Africa or (exclusive) the levantine Asian outskirt, Arabic has spread in all four directions. Westwards, reaching all the way to the Atlantic ocean, northwards through Spain and well into France, southwards into the sub-saharian territories of east Africa, and eastwards well into Uzbekistan and Tajikistan. In the course of History, some

colloquial varieties have gone extinct, as in Sicily or Andalusia, while others have been incorporated by the local languages and disappeared as autonomous languages, like in Malta or Iran.

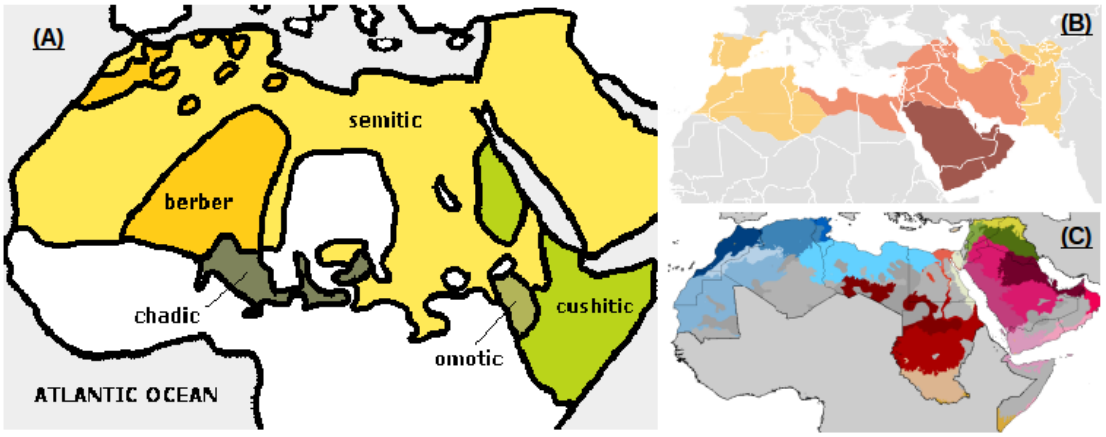

Figure 1.
(A) Map of the Afro-Asiatic language family and its (putative) surviving branches: Semitic, Berber, Cushitic, Chadic and Omotic. With more than 550 million speakers, Afroasiatic is the 4th largest language family in the world and comprises some 375 languages (Ethnologue, 2005). Arabic has spread from north Arabia in all directions with the Islamic conquests **(B)**, and is today the 4th most spoken language in the world, with more than 450 million speakers, exhibiting on its own a similar geographic extension and range of inner diversity as Afroasiatic **(C)**.

As one of the most geographically widespread languages of the world (Behnstedt, 2013) and with more than 450 million speakers, it is not one but many Arabics that are being spoken today, and that form the largest group of extant Semitic languages (Gordon, 2005). Despite a rather shallow origin that can be dated back to the death of the Prophet Muhammad in 632 C.E, Modern Arabic varieties display a significant range of differentiation in all fields - phonology, morphology, syntax and lexicon - to the extent of mutual unintelligibility between geographically distant forms. But this diversity not only is bounded - within a recognizable 'Arabic' unit - but is also structured. It has been shown for instance that subjects were sensitive to the rhythmic

structure of Arabic dialects and relied on prosodic cues to distinguish them, with the most salient distinction being between levantine and north-African dialects - 97 % efficiency for Arabic-speakers, dropping to 56 % for non-Arabic speakers (Barkat et al., 1999). Subsequent studies (Hamdi et al., 2004) demonstrated the existence of a geographically-structured east-to-west perceptual continuum of Arabic dialects based on syllabic structure/complexity – i.e. on the specific alternation patterns of consonants and vowels within words, as well as a transitional zone comprising Tunisian and Egyptian Arabics that display mixed levantine and north-African characteristics.

This remarkable diversity within recognizable linguistic boundaries is the result of the natural evolution of any language - no language is homogeneous – that can ultimately lead to a marked language speciation. Part of this evolution is triggered by language-internal factors within speech communities, while others are linked to these communities own demic history. In the case of Arabic, the Islamic conquests of the 7th to 9th centuries caused a rapid and directional spread of Arabic-speaking tribes from their northern Arabian cradle, with several linguistic consequences. Entire pre-islamic populations abandoned their language to shift to the more prestigious language of the invaders, and as the conquest front advanced, isolation-by-distance and socio-linguistic niche adaptation to the preexisting linguistically, culturally and socially differentiated substrates resulted in marked linguistic diversity. At the same time, due to their common religious culture, the newly settled populations kept interacting economically and through a pervading use of standardized linguistic norms in religious and literary activities - and, in the contemporary Arabic World, through News and Education. Such homogenizing processes tempered the distinctive isoglosses of local varieties, melting them down into larger identifiable linguistic units (Palva,1982).

The development of Arabic dialects is thus the outcome of both divergence and convergence processes that are deeply linked to the history of their speakers, but also to the surrounding non-Arabic substrate, typically Berber and the sub-Saharan languages of east-Africa. But while the course of this demic history - especially its early stages - is abundantly documented in historical archives, the concomitant linguistic developments aren't as much, calling for a different outtake on the matter, and more specifically in connection to the (substantial) demic component of their evolution.

Deme and Idiom polarization in Arabic historical dialectology

Early Arab grammarians were focused on preserving the tongue of the Coran - known as Classical Arabic - through methodical descriptions, and Sîbawayh's, albeit written in the 8th century, remains the grammar of reference for Arabic to this day. The prestige attributed to Classical Arabic hindered the study of vernacular dialects considered to be its low, distorted offsprings from a direct linear descent, and their proper linguistic study was only reignited in the 19th century with the presence in the field of European dialectologists and the production of monographies specifically dedicated to their description. But to this day, the development of contemporary dialects is still a matter of debate for Arabicists and Semiticists alike, and several theories have been proposed to account for the differences among these dialects, and with earlier attested varieties.

Two rather exclusive explanatory paradigms emerge from the scholarship of Arabic historical dialectology - *language genetics* and *language contact* - that oppose two scenarios of contemporary Arabic dialects' genesis (Al-Jallad, 2009). The first tends to assume a lineal descent of modern varieties from a single point of origin located back in time, supposedly Classical Arabic - as defined by Sîbawayh, although forms unattested in the accounts of Arab grammarians have also been considered and tentatively reconstructed from the comparison of modern varieties (Owens, 2006). Whatever the chosen origin, this approach assumes that despite contact episodes, the underlying structure of these dialects' development is distinctively genetic, which could be translated in a phylogenetic paradigm as essentially tree-like (Fig. 2.A). The second perspective assumes on the contrary heterogeneity at all levels of these dialects development : from the inception, with a heterogeneous initial substrate – called a *koine* (Ferguson, 1959), then, through the prevailing contacts in their subsequent history. Consequently, this model assumes that polygenesis of Arabic dialect and diffusion of linguistic traits in the geolinguistic space and across dialect boundaries (Fig. 2.B) to be the structuring forces of these dialects' diversity.

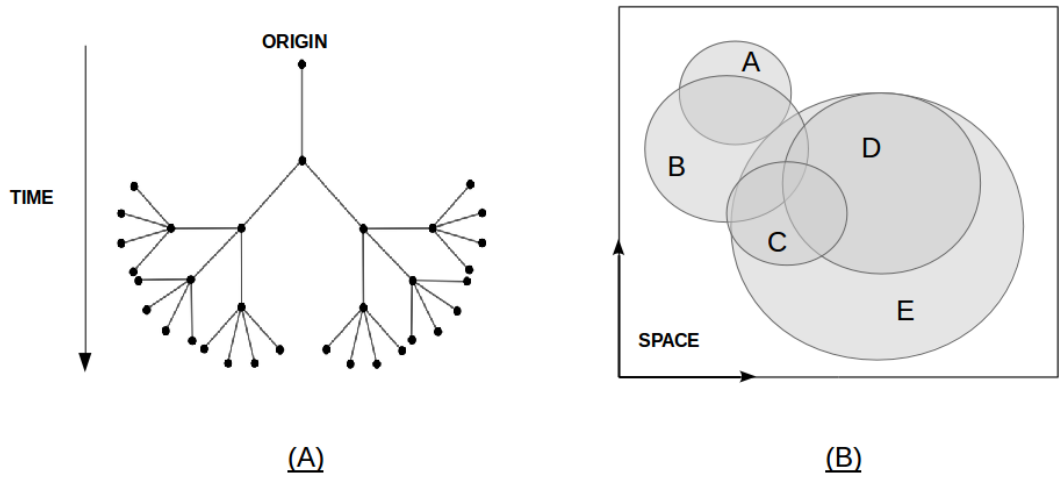

Figure 2.
The tree **(A)** and the wave **(B)** models of language evolution. The tree model assumes that diversification along the axis of time is the essential structuring factor of language diversity, while the wave model assumes that it is the diffusion of linguistic traits/innovations in the geolinguistic space that constitutes its defining factor.

However, polarizing *language genetics* and *language contact*, the tree and the wave, time and space, also polarizes the linguistic and the demic components of language evolution, which are interdependant, decreasing, as Al-Jallad points out, the explanatory power of either perspectives. Since the evolution of Arabic dialects is a mix of such polarities, we need a mixed model of evolution to explore the structure of their relationship in order to capture the various layers of information structuring their variance.

Al-Jallad (2009) submits such a mixed model of evolution for these dialects in order to take into account both the intrinsic heterogeneity of language at each stage of its evolution, and the development in time of identifiable linguistic distributions. His two-step approach first retraces the development of innovations from proto-Central Semitic - regardless of the dialect, then maps the contact-driven diffusion of these features throughout the dialectal domain based on their distribution, with an exemplary application on grammatical features (Fig. 3).

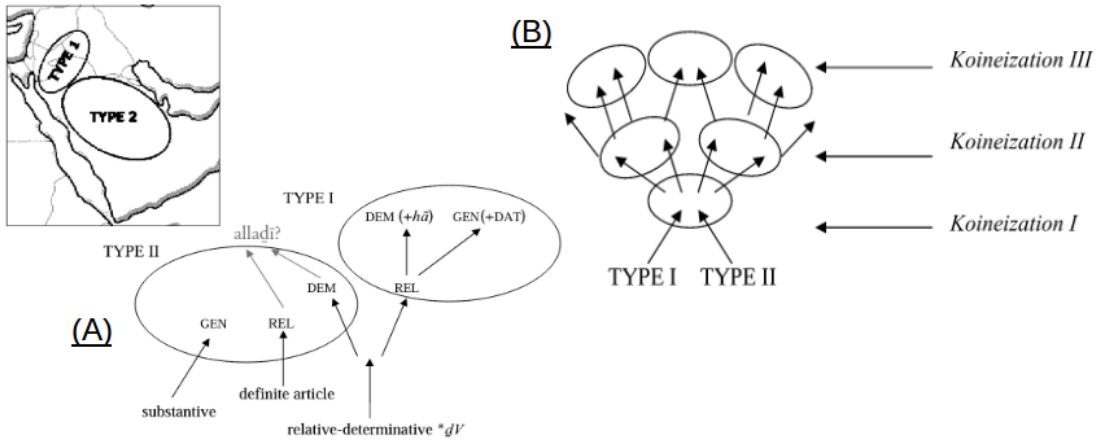

Figure 3.
Al-Jallad (2009)'s model starts by identifying innovations **(A)** through a comparative analysis of grammatical features - REL: relative pronoun; DAT: dative; GEN: genitive particle; DEM: proximal masculine singular demonstrative pronoun - in his language sample vs. Proto-Central Semitic, then traces the diffusion of those innovations in the geolinguistic domain covered by his sample to account for the heterogeneity of their distribution. His aim is to reconstruct putative koineization stages, where each stage corresponds to the production of variants without speciation **(B)**.

With 6 Arabic varieties and 3 grammatical features, Al-Jallad's focus is on establishing the evolutionary trajectory of features (Fig. 3.A) rather than – and somehow in disconnection with – that of the languages. The potential for generalization of this approach beyond this specific case-study is however far from obvious, as is the potential to systematically test the claims it generates against new or different languages or features, typically in terms of the demic history that generated their observed distributions. Moreover, the interaction between genetics and contact in the shaping of variants distribution is not addressed, since this type of analysis considers them sequentially, as if the observed diversity was only a matter of feature recombination, rather than the product of historically-constrained chains of events. Despite these limitations, Al-Jallad's methodological proposal is quite singular in the landscape of Arabic dialectology where studies of vernacular varieties is dominated by descriptions

of individual dialects, atlases and corpus-based socio-linguistic studies (Owens, 2003). Despite a wealth of data, comparative (diachronic) approaches and (synchronic) studies of variation tend to be restricted to one linguistic phenomenon (essentially phonetic or grammatical), one dialect or to the comparison of Arabic as a whole with other languages (see for instance the landscape being sketched for the discipline in Al-Wer and de Jong, 2009).

Words and Language phylogenetics

Inspiration for further historical and comparative analysis of these dialects can however be found outside the field of Arabic dialectology, as the problem of non-independence that arises from historical relatedness is not specific to the case of Arabic, or even to that of dialect and language formation, but encompasses any diversity that was partially generated through lineal descent, whether cultural or biological. In the case of language, a growing interest has been building up around the use of computational phylogenetics method to explore both the shape and the tempo of evolution of linguistic species (Gray and Atkinson, 2003 ; Bouckaert et al., 2012).

The lexicon of a language can be construed as its DNA, in words as genes : they are sequences of sounds that perform specific linguistic functions – different meanings which can be either grammatical (like in Al-Jallad's study) or lexical (as in our case). In the course of language evolution, words are transmitted from one generation to the next with modification, producing at the level of the community of speakers stabilized sound changes in their phonetic sequence. And when population come in contact, they can borrow words from each other and maintain them as their own in the vocabulary. Although the rate of change is higher, language transmission is social and not parental and horizontal transfers are more prevalent in the case of languages, the cumulative effects at the level of the language are similar to what is known for biological species. Hence, it seems appropriate to resort to methods from molecular phylogenetics, applied to lexical comparisons in order to reconstruct patterns of language evolution.

Practically, wordlists – called Swadesh lists - are used to compute such phylolinguistic analyses. Morris Swadesh (1952) introduced a lexicon-based model for language change relying on test-lists of meanings likely to be found in all cultures - such as body parts, lower numerals, topographical terms, kinship terms, personal,

demonstrative and interrogative pronouns, naturally-occurring phenomena and basic human activities. The most common list are 100 and 200 words long, and as putative cultural universals are supposedly less susceptible to change than the rest of a language's lexicon - either naturally or through borrowing, and therefore, can provide a good word sample to reconstruct the history of the languages by maximizing (theoretically) the (phylogenetic) signal to noise ratio.

Closest to Arabic, Kitchen et al. (2009) used such 100-wordlists in computational phylogenetics framework to study the inter-lexical divergence patterns of Semitic languages, in which two Arabic dialects (Moroccan and Somalian (Ogaden)) ensure the monophyly of Arabic (Fig. 4), and puts their divergence data at some 850 of the current era. While their Bayesian approach allows the dating of language divergence and the translation of branch lengths into absolute dates when some external data is available for calibration, it relies on a tree-model of divergence that has been proven problematic in cases of massive contacts between languages and dialect chains. Ben Hamed (2005) and Ben Hamed and Wang (2006) showed for instance in the case of Sinitic – commonly known as Chinese dialects and which are even more divergent than Arabic dialects appear to be given their longer -5000 year - span of evolution, that the tree model led to spurious topologies, and that the conflicting, non-treelike signals that could be captured by phylogenetic networks were as informative about dialects development as they were about their demic history. More generally, Greenhill et al. (2009) showed through simulations that phylogenetic tree estimates are quite robust to moderate borrowing and moderate undetected borrowing (<20 % per millennium), but become increasingly inaccurate as borrowing rates increase.

In most cases however, we don't have prior estimates of how rampant language admixture has been, and consequently, the amount of conflicting non-tree like signal in the data. As generalizations of the tree-model, phylogenetic networks appear therefore to be better suited for empirical exploratory research (Gray et al., 2010), as they seem to capture both the trees and the waves of evolution – linguistic or cultural, but also biological (Nakhleh, 2010). In the case of language, Neighbor-Nets (Bryant and Moulton, 2004) are often used to relax the (mathematical) constraint of treeness and map out - in location and quantity - the distribution of the residual similarity that cannot be explained by the strict tree model.

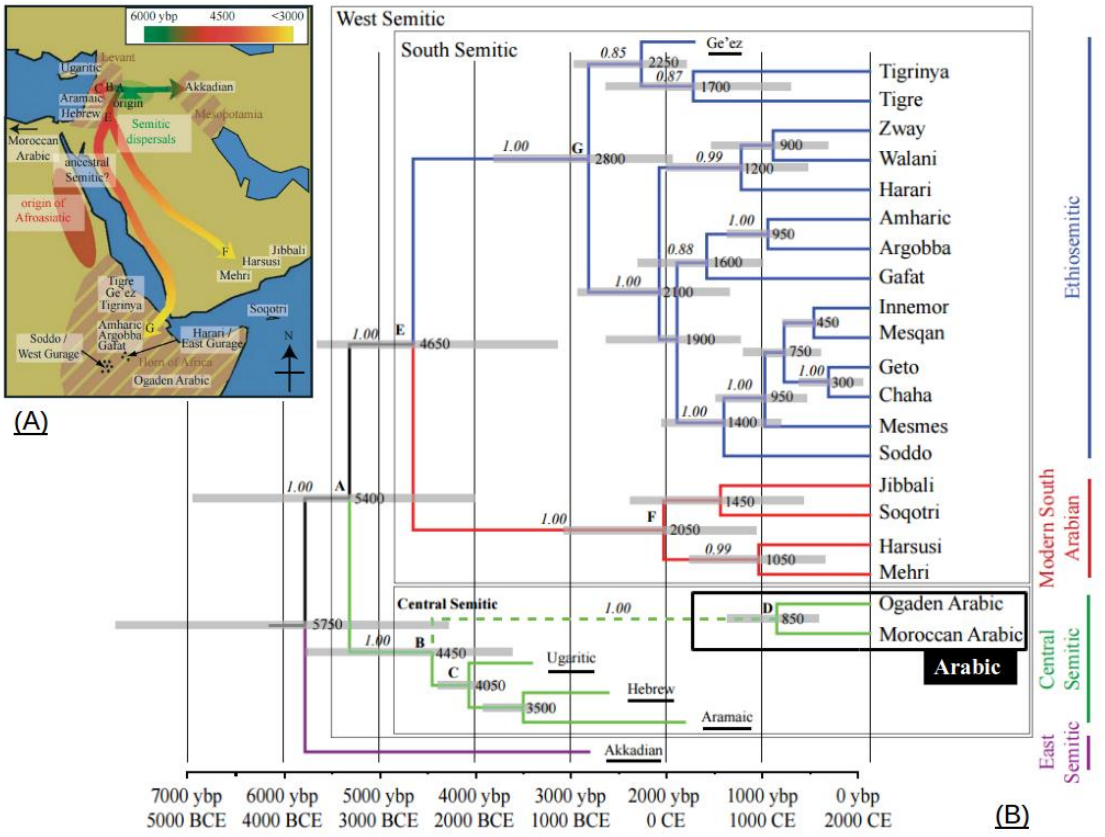

Figure 4.
Bayesian phylogeny of 25 semitic languages reconstructed by Kitchen et al. (2009) from lexical data (100 words). The map in **(A)** shows the putative and attested locations of languages as well as the assumed location of the divergence of ancestral Semitic (in italics) and the dispersal routes (with time scale), inferred from the phylogenetic reconstruction of this language family **(B).** Mean divergence times are indicated to the right of each node, as well as the 95% highest posterior density intervals in the form of light gray bars. The scale bar along the bottom of the phylogeny presents time in YBP, and branch support (posterior probabilities) are indicated in italics above each branch. The tree is rooted with Akkadian. Extinct languages are underlined and subgroups identified by color bars.

This interdependence of lineal (vertical) descent with other (non-vertical) processes of evolution, creates patterns of reticulations between the evolving units : the more un-treelike the signal in the dataset, the more reticulate the resulting picture, providing a principled basis for exploring the underlying evolutionary processes structuring branches and nets. Reticulations show alternative tree-structures supported in the data (Fig. 5) and thus provide alternative explanations for the observed similarities between languages, as well as their relative support (branch length and

confidence). Neighbor-Nets have been applied productively to problems at all levels of the language family tree - higher up to address family memberships (Bryant et al., 2005, for Indo-European), down to the level of dialects (Ben Hamed, 2005, for Chinese dialects) and to address classification issues as well as demic and cultural evolutionary hypotheses.

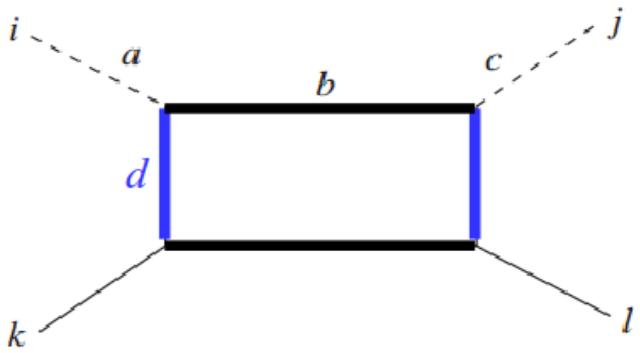

Figure 5.
Splits are the building blocks of Neighbor-nets. The distance between taxa (*i, j*) is the sum (*a*+*b*+*c*). The split in black is the edge identifying a partition between (*i, k*) and (*j,l*), supported by a branch length of *b*, while the split in blue identifies a partition between (*i,j*) and (*k,l*), supported by a branch length of *d*.

Arabic data and dialect inter-lexical distances

Neighbor-nets provide the ideal paradigm to explore the diversity of Arabic dialects in connection with demic processes. Their computation requires a distance matrix of inter-lexical divergence between dialects. To that effect, M. Barkat-Defradas and R. Hamdi-Sultan compiled Swadesh 100-worlists by recording and transcribing native speakers of 12 Arabic dialects (Saoudian, Yemeni, Kowetian, Jordanian, Lebanese, Syrian, Egyptian, Libyan, Tunisian, Algerian, Moroccan and Mauritanian), but also for Chaoui, a Berber language spoken in east Algeria. Rachid Ridouane provided data for Standard Arabic (coded as ArabicST) and for Tachelhiyt, a Berber language spoken in south east Morocco (see for instance Ridouane, 2014). Both Berber languages presented polymorphism - multiple words for one meaning entry, so we recoded them as two entries (Tachelhiyt1/2 and Chaoui1/2 respectively). Stefano Manfredi provided data for the Arabic creole of Juba (see for instance Manfredi, 2013), which is derived from Sudanic Arabic and acts as a major lingua franca of South Sudan,

in both its most prestigious form – the acrolect (JubaAcro) and the basilect (JubaBasi) at the other end of the socio-linguistic continuum. As with Berber, these creoles provide a basis to study the interaction of Arabic with its neighboring non-Arabic languages.

This core data was supplemented by a subset of languages from Kitchen et al. (2009)'s semitic sample, namely 2 extinct languages (Akkadian and ancient Hebrew) and 4 modern south Arabian non-Arabic languages (Harsusi, Mehri, Jibbali and Soqotri – see Fig. 4.A for locations). Hebrew and Akkadian provide ancient material sampled from Central Semitic for the former - of which Arabic is also a member, and for the latter, the earliest attested semitic language and best candidate for rooting Semitic according to Kitchen et al.'s analysis. As for Modern Arabian Languages, they constitute isolated non-Arabic patches in otherwise Arabic-speaking region, and are likely proxies for the pre-Arabic substrate of the region. We did not include more languages to keep the ratio Arabic to non-Arabic even in our largest sample.

The dominant practice in phylolinguistics has been to rely on cognate coding whereby a word in language *A* is said to be cognate with a word in language *B* if their form share some phonetic features - phonetic similarities – that can be interpreted as a sign of common ascendance. Proceeding from comparative Swadesh word lists, words are compared for cognacy across the language set, leading, for a given meaning to identify as many cognate sets as the comparison requires. However, this coding step, usually performed manually by an expert, is only available for some language groups, and Arabic is not one of them. We therefore turned to weighed phonetic alignment (List, 2012) to recode our 100-words Swadesh list in terms of phonetic distance between languages, identify cognates and derive the lexical divergence distance matrix between languages. Phonetic alignment proceeds from the analogy between genes as sequences of nucleotides with words as sequences of sounds (phonemes), by specifying a transition matrix weighing the cost of aligning each two sounds together and computing the optimal global alignment, i.e. the alignment that maximizes the sum of similarities between aligned words. This is not unlike the Levenshtein distance that is commonly used in computational dialectology (Levenshtein, 1966 ; Kondrak, 2002, Heeringa, 2004), which is unweighed, but the introduction of a phonetic weighing schema - in our case Dolgopolsky (1986)'s, has been showed to lead to substantially more accurate phylogenies (Wichmann et al. 2010 ; Jäger, 2013). The resulting matrix is then analyzed to compute the corresponding Neighbor-Nets using the SplitsTree freeware (Huson and Bryant, 2006).

Linguistic interconnections and the demic history of Arabic dialects

Figure 6 shows a Neighbor-Net computed on the inter-lexical divergence matrix from the extended dataset containing 13 Arabic dialects and 12 non-Arabic semitic languages of northern Africa and the Arabian Peninsula. The major partition contrasts Arabic varieties versus all others, with a defined structure for Berber on the one hand, and of Arabian semitic languages – modern and extinct, on the other. With respect to Kitchen et al. (2009)'s analysis however, Arabic does not cluster specifically closer to ancient Hebrew – also classified as a central semitic language, than to Akkadian, and both do, despite their long branch, cluster closer to modern Arabian languages than to Arabic, suggesting that despite the major history of diversification of Arabic dialects is internal, rather than external, in connection to non-Arabic substrates. For Arabic proper, two geolinguistic groups are clearly distinguishable, supporting a partition between Northern African dialects and levantine varieties (in blue) consistent with the perceptual divide put forth by Barkat et al. (1999), but also the transitional nature of Tunisian and Egyptian (Ghazali et al. 2005), albeit in different ways (Hamdi, 2007) : Tunisian is clearly of mixed influence between the two parts of the geolinguistic domain, while Egypt is a transitional post between the Levant and the Juba creoles of South Sudan.

The emergence of Islam in the 7th century unified the Arabian tribal populations that engaged in long-range expansions, starting with the Fertile Crescent and Egypt, this latter acting as the primary base for raids further west towards the Maguire, but also south towards Sudan, Ethiopia and Erithrea. These first conquests by the Umayyades were followed by several migrations, but it is the major flow of the Banu Hilal and Banu Sulaym Arabic tribes during the 11th century that profoundly changed the demic and linguistic makeup of north Africa that was until then only sparsely and scarcely populated by the Berber tribes. The pause that preceded these major migration waves can account for a closer proximity of Egyptian Arabic to both levantine dialects – the longer settlement allowing for the retention of features from their Arabian origin, but also for the mixed nature of Tunisian, which was the first settlement step in the conquest of the region (Abun-Nasr, 1987).

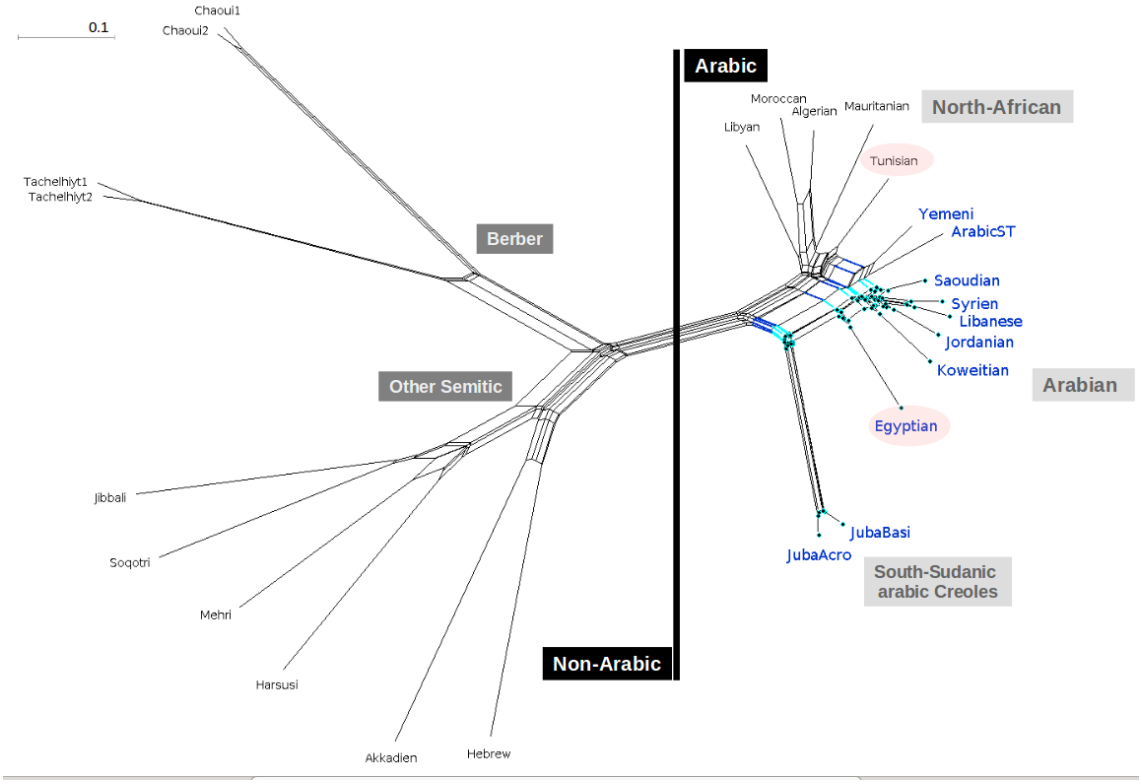

Figure 6.
Neighbor-Net of inter-lexical divergence for 25 languages : 13 Arabic dialects and 2 varieties of Arabic creoles (Juba), 4 Berber varieties and 6 other semitic languages located in the Arabian peninsula (see map in Fig. 4.A), among which Akkadian, which was best supported in Kitchens et al. (2009) as the first divergence from ancestral semitic. Arabic appears clearly monophyletic, and within Arabic, a north-African vs. Arabian (in blue), with the Juba creoles clustering closer to the Arabian sample (nodes and edges in cyan), and more specifically to the Egyptian dialect. The transitional character of Tunisian submitted on rhythmic grounds by Hamdi et al. (2004) is supported but not that of Egyptian.

As for Berber, assimilation was total in the pastoral parts due to the similar way of life of the Arabic invaders. The islamization of Berber tribes played an essential role in their arabization, but the (Afroasiatic) genetic proximity between Arabic and Berber languages was also a primordial factor for this linguistic swipe. Because this language shift was produced by adults (not children language acquisition), their original language left indelible traces in the vernacular varieties of Arabic they came to speak that can account for the interconnections between Berber languages and north-African Arabic dialects (nodes in light blue) in Figure 7.

For the Juba creoles, their singular position in the otherwise clear dialect continuum of Arabic fits with the statement of Manfredi and Tosco (2014) that although these creoles cannot be strictly considered as Arabic dialects, they bear the strong influence of Arabic through its Sudanic dialect.

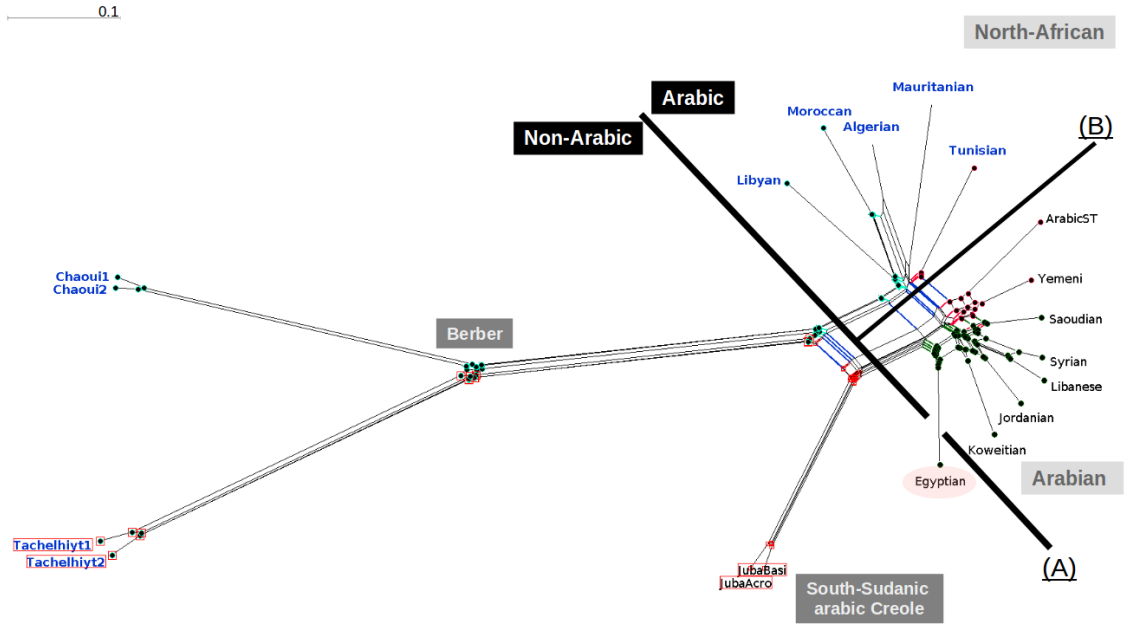

Figure 7.
Neighbor-Net of 13 Arabic dialects and 6 non-Arabic geographically-close languages. **(A)** is the major split in our dataset that clearly identifies an Arabic and anon-Arabic group, and **(B)** the second major split that identifies north-African (in blue) from Arabian languages/dialects, with Egyptian (circled) as a transitional zone between Arabia and the sub-Saharan varieties of Juba. The reticulations between Arabic varieties mark a geographically structured dialect continuum, with transitional positions for both Tunisian (in red – all nodes in red correspond to the languages involved in the conflict/reticulation) and Egyptian (in green). In light blue are the nodes involved in the reticulations across north-African languages, with remnant cross-influences between Moroccan and Libyan Aarabics with the Chaoui and Tachelhiyt Berber varieties. Squared in red are the cross-influences between Berber and Juba varieties.

This is confirmed by Figure 7, when the influence of Kitchen et al. (2009)'s exogenous data is removed and we revert to our core data comprised of Arabic, its creoles and Berber. On this dataset, the major partition pits the Arabic dialect continuum (proper) against Berber and the Juba creoles, which does not invalidate the presence of Arabic influence from Egyptian and levantine varieties, as showed in light blue on Figure 6. It is likely that the prolonged co-existence with Sudanese Arabic has produced increasing structural affinity between them, and although that 'Arabic interference' with the Nilo-Saharian substrate varied according to socio-linguistic variables, it percolated both in the higher acrolectal variety spoken by urban and (Arabic) educated people, and in the basilectal varieties spoken in more rural or recently urbanized areas.

Finally, Figure 8 shows a different representation of language interconnection within the same analytical paradigm of phylogenetic networks. Whereas Neighbor-nets jointly represents the tree-like signal and the non-tree-like signal present within each 4-species/languages set, a reticulogram (Makarenkov and Legendre, 2004) is an augmented phylogenetic tree augmented with additional short-cut edges (in blue). Like Neighbor-nets, it is computed from a distance matrix, in two steps : first a phylogenetic tree is constructed using a method such as neighbor-joining, then additional edges are added to the tree in order to optimize the least square fit of the path distances to the ones in the distance matrix. While slightly less efficient at recovering known reticulations than Neighbor-nets (Huson and Scornavacca, 2011), this representation is slightly more accessible, as it only represents significant hybridizations over a referential tree structure. This method is implemented in the T-Rex analytical freeware (Broc and Makarenkov, 2012).

The reticulogram in figure 8 confirms the previous conclusions drawn from the Neighbor-Nets, but also highlights an interesting connection between Tunisian and Yemeni Arabics, consistent with the idea presented by Abun-Nasr (1987) that although it is difficult to trace the tribal composition of the first Muslim armies, Yemeni tribes are likely to have formed the bulk of the contingents that conquered Egypt in the middle of the 7th century, and from which the Banu Hilal migratory waves will then flood north Africa through Tunisia.

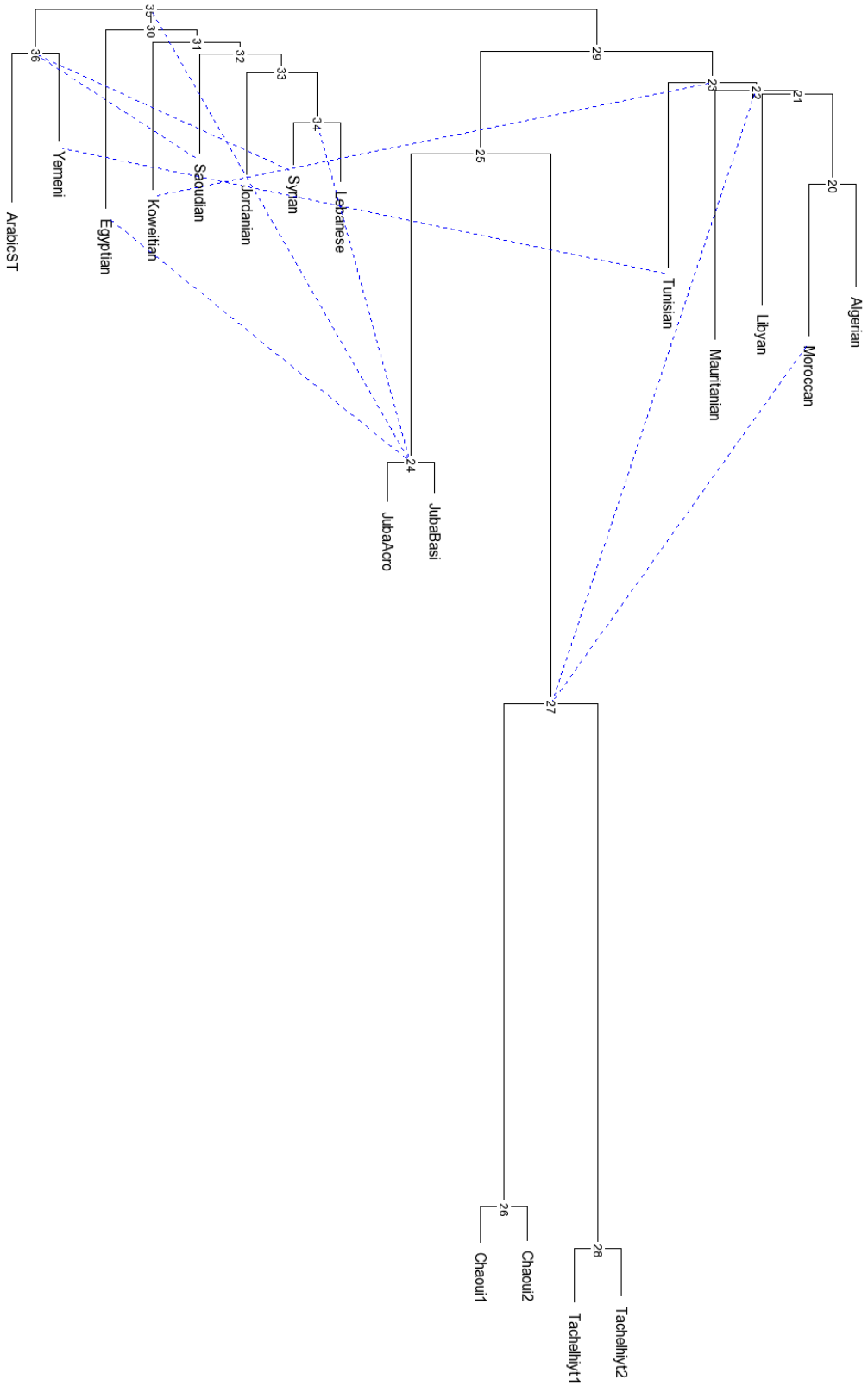

Figure 8.
Reticulogram of the previous sample, in which conflict (dashed lines) is added onto the tree representation. In this case, conflict can be interpreted as contact, and summarizes the information at the level of the node rather than that of the leaf (language).

Fine grained language-based demic history of our species

We hope to have provided here an example of how a comparative linguistic analysis can open a window onto the demic history of these languages' evolution. With its rich inner diversity, Arabic provides a still frame of the natural evolution of language, caught between its internal processes, the demic history of its speakers and the social constraints they impose on it, and this complex history is layered in the material substance of the language – its words. Some similarities between the dialects are the legacy of a distant past while others are the process of local diffusion, and with both meaningful and significant differences *and* similarities between them, the challenge is to analyze their diversity in a way that can capture the synthetic macro-patterns of language interconnection, but also specify their origin – internal/demic/social. Geographic and/or sociological classifications of modern Arabic dialects often rely on phonological units that do not hold up to scrutiny or further sampling (Embarki, 2008), but computational phylogenetics, and specifically phylogenetic networks, offer the means to a systematic exploration of these dialects diversity that can provide the basis for linking their structural makeup to the processes that shaped them.

These powerful representational tools face however a major diagnostic limitation. Whatever the analytical paradigm - tree, wave or network, the challenge is in linking the linguistically-informative micro-level of the datum the analyst can confront to linguistic scholarship, with the process-informing macro-level picture of relationships between languages/dialects, which can be interpreted in evolutionary terms. Being distance-based, and therefore reliant on cumulative similarity between languages/dialects over all data, these networks cannot trace back their steps to the diagnostic micro-level of the datum, and thus link the patterns of treeness and reticulations to the features actually responsible for them, or to the mesoscopic level of the cognate hypothesis they support. They do however provide evolutionary hypotheses that can be tested (externally) through tree-based or geographic (diffusion-based) models, or confronted to external (historical, genetic, archaeological) sources.

The analytical framework that we presented is unlimited in its application, and can be opened up to any dialect sampling, and even linguistic data – lexical, but also

phonological or grammatical. Given the immense diversity of Arabic, we have only captured here a glimpse of its history, that needs to be enriched by further language samples, first through its eastern and sub-Saharan varieties and creoles, but also through finer geo-sociological samples, especially in the Levant, where local varieties may have retained a precious part of the history of both the language and the people that are lost in the typified national varieties. Such links as the one between Koweitian Arabic and north-African varieties is unexpected, and only a richer sampling can help us explore its mystery.

References

Abun-Nasr, J. M., 1987: A history of the Maghrib in the Islamic Period. *Cambridge University Press*, Cambridge.

Al-Jallad A., 2009 : The polygenesis of the neo-arabic dialects, *Journal of semitic studies* 54 : 515–536.

Al-Wer, E., and de Jong, R. (Eds.), 2009 : *Arabic dialectology: in honour of Clive Holes on the occasion of his sixtieth birthday*. Brill.

Barkat, M., Ohala, J., and Pellegrino, F., 1999 : Prosody as a distinctive feature for the discrimination of arabic dialects. *EUROSPEECH* 99 : 395-398.

Behnstedt P. and Woidich M., 2013 : Dialectology. *The Oxford handbook of Arabic linguistics.* Oxford University Press, Oxford, pp. 300–325.

Ben Hamed, M., 2005 : Neighbour-nets portray the Chinese dialect continuum and the linguistic legacy of China's demic history. *Proceedings of the Royal Society of London B: Biological Sciences*, *272*(1567) : 1015-1022.

Ben Hamed, M., and Wang, F., 2006 : Stuck in the forest: trees, networks and Chinese dialects. *Diachronica*, *23*(1) : 29-60.

Ben Hamed, M. and Darlu, P. , 2007 : Gènes et Langues: Une longue histoire commune? *Bulletins et Mémoires de la Société d'Anthropologie de Paris* 19: 243-264.

Ben Hamed, M., 2015 : Phylo-linguistics: Enacting Darwin's Linguistic Image. In *The Handbook of Evolutionary Thinking in the Sciences*, Heams, Huneman, Lecointre & Silberstein (Eds.), pp. 825-852, Springer.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., and Atkinson, Q. D., 2012 : Mapping the origins and expansion of the Indo-European language family. *Science 337*(6097) : 957-960.

Boc, A. and Makarenkov, V., 2012 : T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Research*,*40*(W1) :W573-W579.

Bryant, D., and Moulton, V., 2004 : Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution*,*21*(2) : 255-265.

Bryant, D., Filimon, F., and Gray, R. D., 2005 : Untangling our past: languages, trees, splits and networks. *The evolution of cultural diversity: A phylogenetic approach*, Mace, Hlden and Shennan (Eds.), pp. 67-83. Left Coast Press.

Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. and Mountain, J. , 1988 : Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences 85*(16) : 6002-6006.

Cavalli-Sforza, L. L., Minch, E. and Mountain, J. L., 1992 : Coevolution of genes and languages revisited. *Proceedings of the National Academy of Sciences*, *89*(12) : 5620-5624.

Dolgopolsky, A. B., 1986 : A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia. *Typology, Relationship, and Time: A Collection of Papers on Language Change and Relationship by Soviet Linguists, Shevoroshkin and Markey (Eds.). Karoma, Ann Arbor, MI*, 27-50.

Embarki, M., 2008 : Les dialectes arabes modernes: état et nouvelles perspectives pour la classification géo-sociologique. *Arabica* 55(5) : 583-604.

Ferguson C.A., 1959: The arabic koine, *Language* (2006) : 616–630.

Ghazali, S., Hamdi, R. and Knis, K., 2005 : Intonational and Rhythmic patterns across the Arabic Dialect continuum, *Perspectives on Arabic Linguistics XIX*. Current Issues in Linguistic Theory, Benmamoun (Ed.), pp. 97-121. John Benjamins.

Gordon R. Jr., 2005 : Ethnologue - Languages of the world, 15th edition [http ://www. ethnologue.com].

Gray, R. D., and Atkinson, Q. D., 2003 : Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature 426*(6965) : 435-439.

Gray, R. D., Bryant, D., and Greenhill, S. J., 2010 : On the shape and fabric of human history. *Philosophical Transactions of the Royal Society B: Biological Sciences 365*(1559) : 3923-3933.

Greenhill, S. J., Currie, T. E., and Gray, R. D., 2009 : Does horizontal transmission invalidate cultural phylogenies? *Proceedings of the Royal Society of London B: Biological Sciences*, *276*(1665) : 2299-2306.

Hamdi, R., Barkat, M. & Ben Hamed, M., 2004 : Discrimination des langues et dialectes arabes par le rythme, *Actes de Journées d'Études sur la parole*, Fès, Maroc, 19-22 Avril.

Hamdi, R., 2007 : La variation rythmique dans les dialectes arabes. *Doctoral dissertation*, Université Lyon 2.

Heeringa, W., 2004 : Measuring dialect pronunciation differences using Levenshtein distance. *Doctoral dissertation*, University of Groningen.

Huson, D. H., and Bryant, D., 2006 : Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution 23*(2) : 254-267.

Huson, D. H. and Scornavacca, C., 2011 : A survey of combinatorial methods for phylogenetic networks. *Genome biology and evolution, 3 :* 23-35.

Jäger, G., 2013 : Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change 3*(2) : 245-291.

Kitchen, A., Ehret, C., Assefa, S., and Mulligan, C. J., 2009 : Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proceedings of the Royal Society of London B: Biological Sciences*, *276*(1668) : 2703-2710.

Kondrak, G., 2002 : Algorithms For Language Reconstruction, *Doctoral dissertation*, University of Toronto.

Levenshtein, V. I., 1966: Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8): 707-710.

List, J. M., 2012 : SCA: phonetic alignment based on sound classes. In *New Directions in Logic, Language and Computation*, Lassiter and Slavkovik(Eds.), pp. 32-51. Springer Berlin Heidelberg.

Makarenkov, V. and Legendre, P., 2004 : From a phylogenetic tree to a reticulated network. *Journal of Computational Biology*, *11*(1) : 195-212.

Manfredi, S., 2013 : Juba Arabic corpus. Corpus recorded, transcribed and annotated by Stefano Manfredi, *ANR CorpAfroAs – A Corpus for Afro-Asiatic Languages.*

Manfredi S., Tosco, M., 2014 : Arabic-based Pidgins and Creoles. *Journal of Pidgin and Creole Languages* 29 (2), John Benjamins.

Nakhleh, L., 2010 : Evolutionary phylogenetic networks: models and issues. In *Problem solving handbook in computational biology and bioinformatics*, Heath, Lenwood and Ramakrishnan (Eds.), pp. 125-158. Springer.

Owens J. 2003 : *A linguistic history of Arabic*, Oxford University Press, Oxford.

Palva H. 1982 : Patterns of koineization in modern colloquial arabic, *Acta Orientalia* 43 : 13–32.

Ridouane, R., 2014 : Illustration of the IPA: Tashlhiyt Berber. *Journal of the International Phonetic Association* 44 : 207-221.

Swadesh, M., 1952 : Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American philosophical society*, 452-463.

Wichmann, S., Holman, E. W., Bakker, D., and Brown, C. H., 2010 : Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*, *389* (17) : 3632-3639.