The Notion of Perceptual Distance: The Case of Afro-Asiatic Languages

Julien Meyer, François Pellegrino, Melissa Barkat-Defradas and Fanny Meunier
Laboratoire Dynamique Du Langage, UMR 5596 CNRS/Université Lumière Lyon 2
Lyon, France
Julien.Meyer@etu.univ-lyon2.fr, {Melissa.Barkat ; Francois.Pellegrino}@univ-lyon2.fr, Fanny.Meunier@ish-lyon.cnrs.fr

ABSTRACT

This study investigates the notion of perceptual distance among the Afro-Asiatic family. It is based on a "same/different" task involving French subjects and items from 10 languages or dialects (8 Afro-Asiatic: Amharic, Moroccan Arabic, Jordanian Arabic, Tarifit Berber, Touareg Berber, Hausa, Hebrew, Somali and 2 intruders: Armenian and Turkish). The task was judged very difficult by the subjects. Preliminary results seem to indicate that both segmental and suprasegmental features are involved in the discrimination process, and especially broad phonotactics and syllable structure.

1. INTRODUCTION

Studies on perceptual identification of languages provide an efficient way to tackle with the question of *linguistic distance* among languages and/or dialects. This notion is important since it can shed a different light on the phonological typologies of languages. Many experiments have already focused on different aspects of perceptual identification: subjects (babies or adult listeners [1]), kind of stimuli (natural speech or filtered speech without any segmental information [1], [2], [3]), etc. (see [4] for a review). However, even if these works have assessed the performances of human subjects in language identification tasks, the identification of the salient features used is complex, because of the intricate nature of the stimuli (segmental content, intonational patterns, etc.) and of the socio-linguistic background of the subjects [5], [6]. Another limitation is that in former studies, the languages to identify were often not chosen according to linguistic considerations but to provide a benchmark for comparison with automatic systems [7]. Consequently, the test languages are often reduced to a set of well-described languages such as English, German, French or Japanese. To overpass these limitations, experiments with strict selection of the languages have to be driven. Such experiments are proposed in [8], [9] and [10] resulting in interesting trends, both in terms of linguistic and socio-linguistic strategies.

The study reported here focuses on the comparison of languages of the Afro-Asiatic family at different scales (from regional dialects to inter-linguistic comparison). These languages, spoken roughly in the same area, have been selected in order to avoid the side effect of the geographical origin of the speaker on its physiology [8]. Moreover, they share common properties (e.g. back consonants) that prevent subjects to immediately cluster the languages into distinct broad groups. The aim is to build a **perceptual typology** of these languages based on **segmental** (vocalic and consonantal characteristics) and **supra-segmental** features (rhythmic, prosodic and lexicon structures) revealed by a perception experiment. In the long run, this kind of typology is essential to understand the cognitive representation of language in terms of phonological patterns.

Section 2 describes the languages selected in this study. The experimental paradigm, based on a same/different task is explained in Section 3. The results in term of correct answers are given in Section 4 and a first attempt to evaluate the saliency of the different features and to derive a perceptual map of the languages is detailed in Section 5.

2. LANGUAGES DESCRIPTION

The Afro-Asian linguistic family (AA) is one of the most studied linguistic families, besides the Indo-European. One considers at least five distinct branches (Semitic, Chadic, Cushitic, Berber and the Egyptian Coptic isolate). More recently, Bender [11] postulated the relationship of the Omotic languages with the AA phylum. At present, languages of AA family are spoken by several hundred millions of speakers, and some of them are official languages of many countries (Modern Arabic for instance). Two intruders languages have been joined to the language set in order to provide external comparison criteria.

2.1. The Afro-Asiatic Family

The AA family originated several millenaries ago. Indeed, as early as 3000 Before Present (BP), the Egyptian and Semitic branches were already distinct. Though little is known about the emergence of the different branches, according to Gabriel Camps [12], proto-Berber started to widespread through North Africa eight millenaries BP. The current situation remains complex in terms of number of speakers and even in term of number of distinct languages in each branch, in part because of the fuzzy boundary between language and dialect.

The **Semitic languages** spread over a vast area ranging from Atlantic to Iran and from Mediterranean Sea to Ethiopia. This family is divided into several branches along which one finds Modern Hebrew (northwest Semitic), Arabic (Central South Semitic) and Amharic (South Semitic, mostly spoken in Ethiopia).

The **Berber languages** are currently spoken in ten countries: Morocco, Algeria, Tunisia, Libya, Egypt, Niger, Mali, Burkina-Fasso and Mauritania. The number of speakers is almost impossible to evaluate due to lack of any reliable linguistic census and to the related politic stakes. It seems reasonable to consider that the largest populations dwell in Morocco and Algeria with respectively 30 to 40% and 20 to 25% of the whole Berber speaking population. Touaregs are the third most significant group with ± 900 000 individuals unevenly distributed among the Sahel (Niger and Mali) and the Sahara (Algeria and Libya).

The **Chadic branch** covers approximately 125 languages spoken by 130 million speakers who are distributed among Chad, Niger, Nigeria, Cameroon and the Central African Republic. Hausa is spoken by 80% of the speakers of a Chadic language.

The **Cushitic languages** (about 60 languages) are spoken in Ethiopia, Somalia, Djibouti, in north Sudan and Egypt and even in southern Kenya and Tanzania. Estimations of the number of speakers vary from 14 to 40 millions. Among the Cushitic languages one can quote Bedja (North of Eritrea), which is in contact with Arabic and Nuba (nilotic language); Somali (spoken by about 2.5 millions of individuals living at the edge of the Africa's Horn) and Oromo (spoken by about 9 millions of individuals).

2.2. Intruder languages

The language set as been augmented with two non-AA languages spoken in Minor Asia: Turkish (Altaic language) and Armenian (Indo-European language). Turkish sounds related to Arabic by way of borrowing, but contrary to AA languages, it presents front rounded vowels (as in French, the mother tongue of the subjects). Both Turkish and Armenian feature a velarization mechanism and the glottal fricative /h/. Aspiration and voicing contrasts are phonologically significant in Armenian. These languages provide thus an example of a language typologically related to some of the AA languages but still with "non-AA" characteristics.

2.3 Acoustic-Phonetic description of the languages

Beside the two intruder languages, 8 languages/dialects have been selected from the AA family in order to represent different scales of genetic proximity (from the continuum of regional dialects to the inter branch comparison). See Table 1 (on last page) for a summarized description of the ten languages therefore involved in this study.

3. EXPERIMENTAL DESIGN

3.1. Audio Material

An acoustic database of ten languages has been gathered. It consists mainly of read speech or spontaneous translation of the text *The North Wind and the Sun*. For each language four male speakers have been recorded, digitalised (22kHz, 16 bits) and normalised. The extracts have been dispatched as follow:

- two speakers for learning: 2 excerpts (median duration : 3.6 s; 5th/95th percentiles: 2.8-4.5) per speaker;
- two speakers for test: 10 excerpts (median duration : 2.3 s; 5th/95th percentiles: 2.0-2.5) per speaker.

3.2. Paradigm

18 French native speakers, novice in AA languages, volunteered to participate to the experiment. They were formerly asked to familiarize with the ten AA languages thanks to a visual interface programmed in Flash©. During this learning task they could listen to each item as many times as they wanted. Then they performed a same/different language task: 100 pairs of stimuli were presented to them in an order chosen to avoid psychological expedients (3 different orders – no significant differences among the lists, two different speakers in the same language pair, etc.). Reaction times and performances have been recorded. The subjects generally judged the task as very difficult.

4. RESULTS

4.1 Same Language Pairs

Among the ten languages, Hebrew, Moroccan, Somali and Turkish were the only ones for which the answers were significantly higher than chance when presented in pair (see Table 2). It confirms the difficulty of the task and it shows that the presence of perceptually salient characteristics spread on all languages (e.g. back consonants) is very confusing. In France, the subjects may have been previously exposed to Hebrew and Moroccan, resulting in a better *a priori* knowledge of these languages. However, the good scores reached with Somali and Turkish seem to be due to intrinsic characteristics. The very low score for Hausa pair (though not significantly difference from chance) is surprising, since Hausa is the only tone language of this study. However, it is possible that the dialectal differences of the speakers confused the subjects.

4.2 Different Language Pairs

Subjects discriminate significantly 30 of the 45 different languages pairs (see Table 2). The best score is reached with Hebrew-Turkish (85.5 % of correct answer). Hebrew, Somali and Turkish are the most individualized languages while Touareg and Jordanian are the most confused languages (44.1 % of correct answer).

The phonetic labelling of the excerpts (in broad phonetic classes) is in progress and may reveal correlations between the segmental compositions of the items and the identification results.

5. INTERPRETATION

5.1 Data selection

Multidimensional scaling is a convenient way to represent distances, but it is very sensitive to perturbations. For this reason, only the subjects reaching a better-than-chance score on the same language pairs have been considered (9 out from 18). Moreover, considering the complexity of the task, the number of represented languages has been reduced and Armenian and Hausa have been left out. Armenian because Turkish provides a better extern root of the representation, and Hausa because the very low score reached in same language pair means probably that the dialects spoken by the 2 speakers were to different.

5.2 Multidimensional scaling

A 3D PROXSCAL analysis (explaining 97% of the dispersion) is performed [13]. Hypotheses explaining the plot (see Figure 1) may be proposed. Along Dimension 1, a cluster draws together Jordanian and Touareg (the 2 languages for which consonant clusters are forbidden) while Hebrew which is the only language lacking gemination is isolated. Dimension 2 may separate the plane according to the absence or presence of affricates in the languages. Along dimension 3, the fact that Somali is isolated may be a consequence of the VV sequences common in this language even *into* syllables which is unique in the considered languages.

6. CONCLUSION AND PERSPECTIVES

The experiment reported here provides more indications than evidences to deal with the question of perceptual distance and saliency of the features. Preliminary results seem to show that the same/different decision involves both segmental and supra-segmental features and especially those related to the phonotactics and syllable rules (consonantal clusters, affricates, VV sequences, etc.). This first experiment will be pursued with 1. the phonetic labelling of the excerpts; 2. an identification task where subjects will have to explicitly identify the languages; 3. experiments putting the emphasis on rhythm and intonational characteristics since the current experiment does not give much indications on those prosodic features.

AKNOWLEDGEMENTS

The authors would like to thank Naima Louali and Mehmet-Ali Akinci respectively for the Berber and Turkish data; Ian Maddieson for his help with the selection of languages and finally Ivan Magrin-Chagnolleau for his help with the data preprocessing.

This research is supported by the French CNRS (APN no12374 from the Humanities Department) and the French *Ministère de la Recherche* (ACI no 67280).

REFERENCES

[1] F. Ramus and J. Mehler, “Language identification with suprasegmental cues: A study based on speech resynthesis”, *Journal of Acoustical Society of America*, Vol. 105, N°1, p.512-521, 1999.

[2] K. Atkinson, “Language identification from non segmental cues”, *Journal of the Acoustical Society of America*, vol. 44, n°378 (A), 1968.

[3] J.J. Ohala and B. Gilbert, “On listeners’ ability to identify languages by their prosody”, in *Problèmes de prosodie*, vol. 2, Léon & Rossi (Eds), pp. 123-131, 1979.

[4] M. Barkat and I. Vasilescu, “From perceptual designs to linguistic Typology and Automatic Language Identification” in *Proc. of Eurospeech’01 Scandinavia*, Aalborg, 2001.

[5] E. A. Marks, Z.S. Bond and V. Stockmal, “The effect of proficiency in a specific foreign language on the ability to identify a novel foreign language”, in *Proc. Of International Congress of Phonetic Sciences ICPhS’1999*, San Francisco, 1999.

[6] M. Lorch, P. Meara, “How people listen to languages they don’t know”, *Language Sciences*, Vol. 11, N° 4, p.343-353, 1989.

[7] Y.K. Muthusamy, N. Jain and R.A. Cole, “Perceptual benchmarks for automatic language identification”, in Proc. *of IEEE ICASSP’94*, Adelaide, 1994.

[8] V. Stockmal, D. Muljani and Z.S. Bond, “Perceptual Features of Unknown Foreign Languages as Revealed by Multi-dimensional Scaling”, in *Proc. of International Conference of Spoken Language Processing ICSLP’1996*, Philadelphia, 1996.

[9] I. Vasilescu, F. Pellegrino and J-.M. Hombert, “Perceptual features for the identification of Romance Languages”, in *Proc. of International Conference of Spoken Language Processing ICSLP’2000*, Beijing, 2000.

[10] M. Barkat, J. Ohala and F. Pellegrino, “Prosody as a Distinctive Feature for the Recognition of Arabic Dialects” in *Proc. of Eurospeech’99*, Budapest, pp. 395-398, 1999.

[11] M.L. Bender, *Omotic: a new Afroasiatic language family*,. Carbondale: University Museum, Southern Illinois University, 1975

[12] G. Camps, *Les Berbères, aux marges de l'Histoire*, Edition des Hespérides, 1980

[13] J.J. Commandeur and W.J. Heiser, *Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data matrices*, Tech. Rep. No. RR-93-03, Leiden University, The Netherlands, 1993.

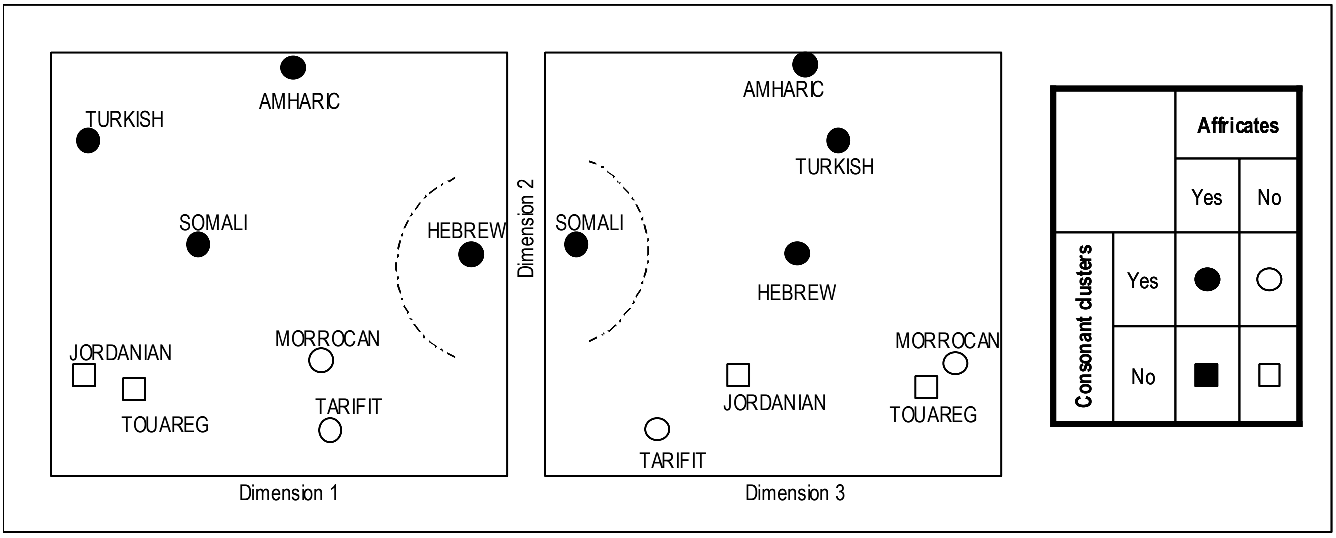
[14] A.S. Kaye and P.T. Daniels, *Phonologies of Asia and Africa*, vol. 1 and 2, Eisenbrauns, Indiana, 1041 p. 1997

	AH	**AR**	**HA**	**HE**	**AJ**	**AM**	**SO**	**TA**	**TO**	**TU**
AH = Amharic	61	54	45	**69**	**72**	**67**	**78**	70	**73**	**67**
AR = Armenian	54	69	55	**67**	62.5	**65**	**74**	49	**72**	**74**
HA = Hausa	45	55	39	**72**	51	**65**	**72**	52	**63**	61
HE = Hebrew	**69**	**67**	**72**	**89**	**65**	**67**	**80**	**71.5**	**82**	**85.5**
AJ = Jordanian Arabic	**72**	62.5	51	**65**	72	62.5	**65.5**	59.5	44	**63**
AM = Morrocan Arabic	**67**	**65**	**65**	**67**	62.5	**83**	**78**	55	61	**75**
SO = Somali	**78**	**74**	**72**	**80**	**65.5**	**78**	**94**	58	**75**	**84**
TA = Tarifit Berber	**70**	49	52	**71.5**	59.5	55	58	67	**66**	59.5
TO = Touareg Berber	**73**	**72**	**63**	**82**	44	61	**75**	**66**	72	**64**
TU = Turkish	**67**	**74**	61	**85.5**	**63**	**75**	**84**	59.5	**64**	**97**

Table 2 – Correct answers (in percent) for each language pair. Matrix is symmetrical.
Scores significantly better than chance are in bold.

Family/Branch		*Language*	*Consonant features*	*Vowel features (phonetic)*	Suprasegmentals
ALTAIC		Turkish	Velarization /l/ [l^{ɣ}] Palatalization Affricates	Front rounded vowels - Vowel harmony - Duration contrast in loanwords	Consonantal clusters allowed
INDO-EUROPEAN		Armenian	Aspiration Velarization Affricates	7 vowels and 4 diphthongs	Consonantal clusters allowed
AFRO-ASIATIC	SEMITIC	Moroccan Arabic	Pharyngeal consonants	12 phonetic vowels i ɛ æ a ɑ u o ʊ ɪ ə iː uː (former diphthongs)	Consonantal clusters allowed Vowel sequences forbidden
		Jordanian Arabic	Pharyngeal consonants	20 vowels Duration contrast i a u ɛ æ e ɑ ʊ ɪ o iː uː aː eː æː eː ɑː oː	No Consonantal clusters CV Vowel sequences forbidden
		Modern Hebrew	Affricateṣ in loanwords No gemination	6 vowels i e ɛ a o u	Consonantal clusters allowed
		Amharic	Ejectives Labialization	7 vowels i e ə u ɐ ɔ a Duration contrast (uncommon)	Consonantal clusters allowed Vowel sequences forbidden
	BERBERE	Tarifit (*temsammane*)	Pharyngealization Velarization	4 vowels and 2 diphthongs i e u a ɛa ɔɑ	Consonantal clusters allowed Vowel sequences forbidden
		Touareg	Pharyngealization Velarization Palatalization	i e u o a ɛ ə ɐ Duration contrast (debated)	No Consonantal clusters
	CHADIC	Hausa	Labialised Velars Palatalized Velars Ejectives Glottalization Retroflexed (ɽ)	- 10 vowels - 2 diphthongs - Duration contrast ɪ ɛ ɑ o u iː eː aː oː uː aj aw	No Consonantal clusters CV / CVV / CVC Tone Language
	CUSHITIC	Somali	Aspiration	- 20 vowels - +/- ATR - Duration contrast - ATR short i e æ u o + ATR short ɪ ɛ a (ʌ) ʊ ɔ - ATR long iː eː æː uː oː + ATR long ɪː ɛː aː ʊː ɔː	Consonantal clusters allowed at syllable boundaries VV intra-syllabic sequence

Table 1 – Potential salient features for identification of the 10 languages/dialects (adapted from [14]).



		Affricates	
		Yes	No
Consonant clusters	Yes	●	○
	No	■	□

Figure 1 – 3D Multidimensional scaling of 8 languages. Shapes and colours indicate respectively presence of Consonant clusters and Affricates. Hebrew is the only language lacking gemination and Somali is the only language allowing VV intra-syllabic sequences.