

Discrimination des langues et des dialectes arabes par le rythme

Rym Hamdi, Melissa Barkat-Defradas, Mahé Ben Hamed

Laboratoire Dynamique Du Langage (UMR CNRS 5596)

Université de Lyon-2, France

Tél.: ++33 (0)4 72 72 65 37 - Fax: ++33 (0)4 72 72 65 90

Mél: rim.hamdi@etu.univ-lyon2.fr, melissa.barkat@univ-lyon2.fr, Mahe.Ben-Hamed@ish-lyon.cnrs.fr

ABSTRACT

This paper investigates the validity of the notion of rhythmic classes, as opposed to the notion of continuum. Our case study concerns six Arabic dialects described as stress-timed. In addition to the comparison among dialects, we also compared them with other languages, some of which are classified in different rhythmic categories. Our approach is based on the computation of the proportion of vocalic intervals and on the standard deviation of consonantal intervals, an acoustic investigation put forth by Ramus & al. [1]. Results show that syllabic complexity and vowel reduction are the main factors responsible for the rhythmic variation of Arabic dialects. However, despite this variation, these dialects continue to cluster around stress-timed languages and to exhibit a different distribution from other rhythm categories.

1. INTRODUCTION

Il existe dans la littérature linguistique un grand nombre de descriptions qui ont porté sur la discrimination des dialectes arabes mais le peu d'études qui se sont intéressées à la variation prosodique de ces dialectes se sont basés sur des épreuves qui étaient essentiellement d'ordre perceptuel [2].

Rappelons que dans cette expérience perceptuelle basée sur de la parole synthétique, des sujets naïfs, natifs de différentes régions du monde arabe, ont été capables de distinguer avec succès les locuteurs issus du Maghreb de ceux issus du Moyen-Orient dans 97% des cas. Pour la même tâche, une population de sujets non-arabophones ont pu distinguer les deux régions dans 56% des cas. Les résultats étant statistiquement significatifs ($p < 0.05$).

Dans une recherche précédente, Ghazali, Hamdi et Barkat [3], ont démontré en utilisant la même méthode que [1] que les dialectes arabes généralement décrits comme 'accentuels' présentent en réalité des structures rythmiques différenciées qui varient géographiquement du Maghreb au Moyen-Orient

Les résultats de cette recherche étaient compatibles avec la notion du continuum rythmique des dialectes arabes, et permettent la simulation d'expériences de discrimination de langues chez des sujets humains. En effet les expériences présentées renforcent la cohérence globale du

modèle proposé par [1]. Des applications à l'identification automatique des langues sont envisagées.

2. APPROCHE PHONOLOGIQUE DU RYTHME

Les langues du monde ont été classées par les linguistes selon leurs propriétés rythmiques et les recherches publiées sur le rythme de la parole en arabe ont classé tous les dialectes étudiés comme étant accentuels (stress-timed) par opposition à syllabiques (syllable-timed) ou à moraiques (mora-timed) (Abercrombie[4]; Miller[5]; Benguerel[6]; Tajima, Zawayda & Kitabara[7]).

De nombreuses études se sont basées sur la théorie de l'isochronie, et suivant cette hypothèse, il serait possible, en mesurant les durées séparant les accents, les syllabes ou les mores dans un échantillon d'une langue, de déterminer la classe rythmique de celle-ci. Malgré de nombreuses recherches, cette hypothèse n'a pas été validée empiriquement.

Les mesures obtenues à partir du signal acoustique ne peuvent pas constituer une preuve suffisante pour classer des langues (y compris l'arabe) comme étant accentuelles par opposition à syllabiques Roach [8], dans l'hypothèse où "stress-timing" signifie bien isochronie des accents.

Dauer [9, 10] qui est parvenue à des conclusions similaires, explique la perception des différents types de rythme essentiellement par les différences de structure syllabique, de réduction vocalique et de type d'accent. Ramus & al. [1] soutiennent que les classes de rythme perçues, et correspondant aux propriétés phonologiques mentionnées par [9, 10], peuvent être mesurées au niveau du signal acoustique. [1] proposent une approche qui repose sur une conception du rythme de parole, non plus comme primitive de l'organisation temporelle des langues, mais comme conséquence de certaines de leurs propriétés phonologiques [8, 9]. Ces propriétés sont notamment : la complexité des syllabes, la corrélation entre poids syllabique et accent, la présence ou non de réduction vocalique des corrélats acoustiques pour mesurer le degré de la complexité syllabique et la réduction vocalique dans chaque langue. Il s'agit d'implémenter le modèle phonologique du rythme à l'aide d'une segmentation phonétique, et de calculer les proportions moyennes des durées d'intervalles vocaliques (%V) et l'écart type des durées d'intervalles consonantiques (ΔC) qu'ils définissent comme ceci :

- Les intervalles vocaliques : toute séquence

ininterrompue de voyelles (du début de la première voyelle de la séquence à la fin de la dernière).

- Les intervalles consonantiques : toute séquence ininterrompue de consonnes.

3. TYPOLOGIE RYTHMIQUE DES DIALECTES ARABES

3.1. Méthode

Corpus et sujets

Dans cette étude nous avons examiné la production du rythme dans six dialectes arabes : Marocain, Algérien et Tunisien représentant l'arabe maghrébin (MA) et Jordanien, Syrien et Egyptien représentant le Moyen-Orient (MO). En outre trois autres langues ont été rajoutées a fin de multiplier les points de comparaison.

Ces trois langues sont classifiées par les linguistes comme appartenant à des catégories différentes : l'anglais, langue accentuelle (stress-timed), le français, langue syllabique (syllable-timed) et le Catalan, langue intermédiaire, trois sujets pour chaque langue et/ou dialecte.

Rappelons que l'expérience nous permettra de voir si les dialectes arabes continuent à être regroupés en tant que langue accentuelle ou s'ils seraient redistribués dans d'autres catégories rythmiques.

Le corpus consiste en enregistrements de l'histoire de "*la bise et le soleil*". Chaque sujet écoutait le texte en français, phrase par phrase, avant de traduire chaque phrase spontanément dans son dialecte ou sa langue d'origine. Pour les sujets français, les phrases ont été présentées en anglais. Les extraits ont ensuite été digitalisés à 22KH/16bits/mono. Le corpus final se compose de 270 phrases (10 phrases/sujet) d'une durée moyenne de 2.5 secondes par phrase.

Analyses

Suivant les procédures expérimentales proposées par [1], nous avons utilisé le logiciel StkCV développé par Pellegrino [11] pour identifier et classer les segments en tant que consonne ou voyelle en calculant :

- (i) La durée de chaque phrase
- (ii) La durée de chaque séquence de voyelles (i.e. intervalles vocaliques)
- (iii) La durée de chaque séquence de consonnes consécutives (i.e. intervalles consonantiques).

Nous avons ensuite mesuré la proportion d'intervalles vocaliques et consonantiques (respectivement %V et %C) dans chaque phrase, et l'écart type des intervalles vocaliques et consonantiques à l'intérieur de chaque phrase (ΔC & ΔV respectivement). [1] proposent que la variable %V est en corrélation avec le phénomène de la réduction vocalique et ΔC est en corrélation avec et la complexité syllabique : les deux propriétés phonologiques qui capturent l'essentiel des propriétés rythmiques des

langues.

3.2. Résultats

Table 1 : Valeurs de %V et ΔC des langues et des dialectes étudiés

Aire dialectale	Langue/Dialecte	%V	ΔC
Maghreb	Marocain	30.81	102.22
	Algérien	31.14	57.54
	Tunisien	34.86	57.54
Moyen-Orient	Egyptien	36.82	52.38
	Jordanien	40.48	55.06
	Syrien	39.76	51.10
	Français	43.38	48.72
	Anglais	38.07	75.27
	Catalan	38.00	60.37

Ces résultats montrent que bien que la proportion d'intervalles vocaliques représente moins de 50% de la durée totale d'une phrase dans tous les dialectes, elle est plus importante dans les dialectes du Moyen-Orient que dans le Maghreb. En fait il y a une augmentation graduelle de %V et une baisse de ΔC au fur et à mesure qu'on se déplace de l'Ouest vers l'Est.

La figure 1 montre la corrélation négative entre % V et ΔC et la position géographique du dialecte ($r = - 0,75$). Quand on va de l'Ouest vers l'Est, ΔC diminue et %V augmente. Des t-test ont ensuite été réalisés pour étudier statistiquement les écarts observés entre paires de dialectes différents. Il apparaît que le seuil de signification est directement proportionnel à la distance entre les deux régions où ces dialectes sont parlés. Ainsi, quand on compare la Syrie à la Jordanie ou le Maroc à l'Algérie, c'est-à-dire deux dialectes issus de la même région, les résultats ne sont pas significatifs alors que pour des paires de dialectes situés aux extrémités opposées du continuum, comme c'est le cas du Maroc et de la Syrie, ils s'avèrent hautement significatifs ($p < 0.001$).

Cependant, il est à noter que l'anglais, le français et le catalan situés dans le continuum montrent des résultats significatifs lorsqu'ils sont comparés à la totalité des dialectes arabes, un fait que nous commenterons plus tard (figure2).

La figure 3 illustre les valeurs moyennes de la proportion des intervalles vocaliques et de l'écart type des intervalles consonantiques, quand les 3 dialectes de chaque région sont regroupés ensemble. Elle montre clairement que %V est plus élevée dans les dialectes du Moyen-Orient (MO) et le français (langue syllabique) que dans ceux du Maghreb (MA) ($p < 0.0001$) qui présentent des valeurs proches de l'anglais et le catalan, alors que l'inverse est vrai pour ΔC .

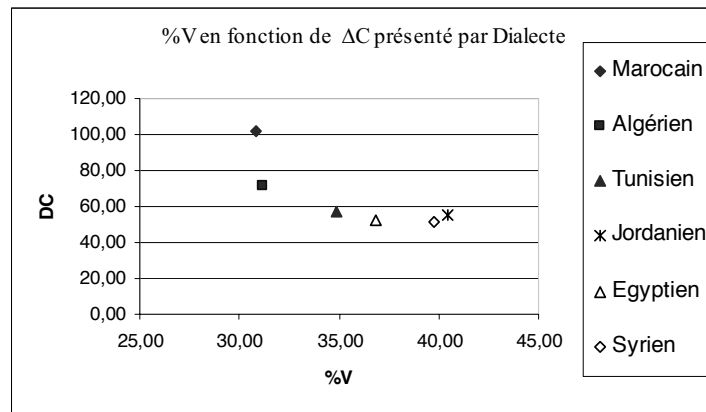


Figure 1 : Corrélation %V/ ΔC dans les dialectes arabes

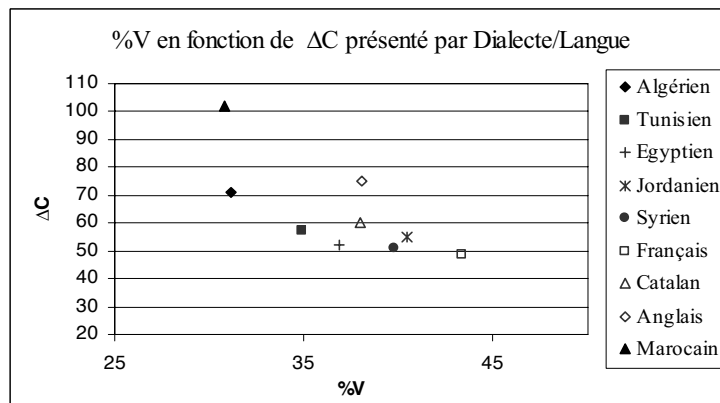


Figure 2 : Distribution des dialectes arabes et des langues sur le plan %V/ ΔC.

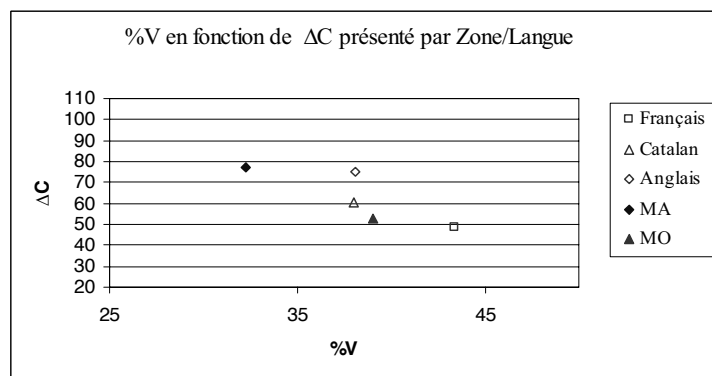


Figure 3 : Corrélation %V/ ΔC dans les dialectes arabes regroupés par zone et les autres langues.

3.3. Discussion

Il est clair que les variables %V ΔC sont directement liés à la structure syllabique. Les syllabes complexes comportent plus de consonnes que les syllabes simples et

Les proportions d'intervalles vocaliques plus basses reflètent la présence de voyelles plus brèves. Toutes les études de l'espace vocalique arabe ont souligné le fait que dans les dialectes maghrébins les voyelles phonologiquement longues ou brèves ont une durée

moindre que les voyelles correspondantes dans les dialectes moyen-orientaux [2]. Il y a une corrélation positive entre l'écart type des intervalles consonantiques et la diversité et la complexité syllabique [1].

L'arabe maghrébin est bien connu pour le processus d'effacement des voyelles brèves en syllabes ouvertes, résultant en différents groupes consonantiques et types de syllabes avec des attaques et marges complexes. Les syllabes complexes associées à des voyelles réduites, spécialement dans les dialectes marocains donnent lieu à cette impression de discours saccadé et haché rapporté par les auditeurs de Barkat [2]. Les différences entre durée vocalique et complexité syllabique semblent être les facteurs principaux responsables des différences dans la structure rythmique. En effet dans cette étude on a confirmé les résultats de [1] concernant les langues : l'anglais présente toujours des ΔC plus hauts et des %V plus bas, c'est une langue qui était traditionnellement classée comme une langue typiquement "stress-timed". Dans notre étude les dialectes dont les valeurs sont proches de cette langue sont ceux parlés au Maghreb tandis que les dialectes orientaux sont plutôt proches du français : langue qui se caractérise par une syllabation simple et une prédominance de syllabes ouvertes.

Puisque les dialectes comme ceux de l'Iraq [5] et la Jordanie [6] ont aussi été classés comme accentuels, nous pourrions peut être envisager une grande variation au sein de la classe "stress-timed". Nos résultats concernant les dialectes confirment la notion du continuum rythmique : certains dialectes sont plus stress-timed que d'autres [9, 10]. Pour maintenir une catégorie accentuelle distincte d'autres types de timing, il devrait exister un ou plusieurs facteurs clés dont la présence induit constamment la perception de l'accent. Un tel facteur serait la tendance, dans tous les dialectes arabes, des syllabes longues et lourdes à attirer l'accent. Puisque le poids syllabique varie graduellement dans ces dialectes, nous pourrions avoir l'impression de différentes sous-classes du rythme. Il est aussi intéressant de remarquer que les dialectes situés géographiquement entre les pôles sont aussi intermédiaires aux points de vue traits phonétiques et se rapprochent d'une langue classée comme intermédiaire qui est le catalan (figure 2). En effet, les locuteurs tunisiens ont un %V similaire aux autres pays du Maghreb mais un ΔC plus proche du Moyen-Orient. Autrement dit, leurs voyelles sont légèrement plus longues et moins réduites que celles des marocains et algériens, mais significativement plus brèves que celles des syriens et jordaniens. Cependant, les locuteurs tunisiens ne montrent pas la même complexité syllabique que les autres sujets maghrébins.

3.4. Conclusion

Les mesures utilisées dans ce travail semblent pouvoir rendre compte des classes de rythme décrites dans la littérature, en outre elles confirment la notion du continuum rythmique et non pas de catégories rythmiques absolues.

Une tâche perceptuelle basée sur de la parole re-synthétisée d'un corpus multilingue pourrait être utile pour compléter ce travail.

Enfin, rappelons que ce travail se limite à étudier un seul aspect du rythme qui est la durée et la structure syllabique, mais il nous a permis de confirmer la notion de variation rythmique dans les dialectes arabes.

3.5. Remerciements

Je tiens à remercier François Pellegrino pour sa collaboration ainsi que Salem Ghazali pour son aide et ses conseils.

BIBLIOGRAPHIE

- [1] F. Ramus. & J. Melher: Language identification with suprasegmental cues. A study based on speech resynthesis. *Journal of the Acoustical Society of America*, 105 (1), pages 512-521, 1999.
- [2] M. Barkat: Vers l'identification automatique des dialectes arabes. *Revue langues et linguistique*, vol.7, pages 47-75, 2001.
- [3] S.Ghazali, R.Hamdi, et M. Barkat: Rhythm Variation in Arabic Dialects. *In Proceeding of Speech Prosody 2002*, Aix en Provence avril 2002.
- [4] D. Abercrombie: *Elements of General Phonetics*, Edinburgh University Press, 1967.
- [5] M. Miller: "On the perception of rhythm", *Journal of Phonetics*, 12, pages 75-83, 1984.
- [6] A. Benguerel: Stress-timing vs syllable-timing vs mora-timing: The perception of speech rhythm by native speakers of different languages. *VARIA, Etudes & Travaux* n°3, 1999.
- [7] K. Tajima, B. Zawaydeh, M. Kitabara: A comparative study of speech rhythm in Arabic, English and Japanese, *Proceedings of the XIV ICPHS*, San Francisco, U.S.A., 1999.
- [8] P. Roach: On the distinction between 'stress-timed' and 'syllable-timed' languages. *In Linguistic controversies*, D. Crystal Ed, pages 73-79. London: Edward Arnold, 1982.
- [9] R .M. Dauer: Stress-timing and syllable-timing reanalysed, *Journal of Phonetics* pages 51-69, 1983.
- [10] R .M. Dauer: "Phonetic and phonological components of language rhythm. *Proceedings of the XIth ICPHS*, Tallinn, Estonia, Vol.5, pages 447-450. 1987.
- [11] F. Pellegrino : StkCV, Logiciel développé au laboratoire Dynamique Du Langage, 2001.