Actes des 8èmes RJC Parole
Espace perceptuel de similarité : étude sur 17 langues

Marie Rimbault-Joffard

Laboratoire de Phonétique et Phonologie, ILPGA
19 rue des Bernardins, 75005 Paris, France
Tél. : +33 (0)1 44 32 05 75
Fax : +33 (0)1 44 32 05
Courriel : marie.rimbaultjoffard@yahoo.de

ABSTRACT

The goal of the present study is to devise a means for representing languages in a perceptual similarity space based on their overall sound structures. In experiment 1, French listeners performed a free classification task in which they grouped 17 languages based on their overall similarity as in [Bra07] with English listeners. A similarity matrix of the grouping patterns was then submitted to clustering and multidimensional scaling analyses. In experiment 2, the same group of French listeners sorted the 17 languages in term of their distance to French. Taken together, the results of the two experiments provide the basis for estimating the distance between a given mother tongue and other languages and for understanding the role of the phonological filter.

Keywords: universals and typology, speech perception, language classification

1. INTRODUCTION

Les études sur la perception des langues se limitent dans leur majorité à des segments phonétiques particuliers (systèmes vocaliques, partie des systèmes consonantiques) ainsi qu'à un nombre relativement restreint de langues (en général deux langues, parfois trois voire quatre). Or, il peut être intéressant, avant de réaliser des études acoustiques approfondies pour déterminer les critères qui sous-tendent la perception et la classification perceptive des langues, de savoir comment celle-ci se réalise d'un point de vue global. En effet, on peut supposer que deux langues A et B issues de la même branche phylogénétique (par exemple le catalan et le galicien), partagent des caractéristiques phonétiques issues de leur héritage commun (même si elles ont chacune évolué indépendamment), et qu'à ce titre elles devraient être perçues par un auditeur français comme assez similaires. Ainsi, la cartographie perceptive des langues pourraient refléter, du moins en partie les affiliations génétiques entre langues.

2. EXPERIENCE 1

2.1. Méthode

Les échantillons des 17 langues ont été sélectionnés (Bra07) parmi celles disponibles sur le site internet de l'Association Internationale de Phonétique. La famille indo-européenne est la plus représentée avec 8 langues : catalan, galicien (langues romanes) ; suédois, néerlandais (langues germaniques) ; slovène, croate (langues slaves) ; sindhi, persan (langues indo-iraniennes), puis la famille afro-asiatique : amharique, arabe, hébreu (langues sémitiques) ; haoussa (langue tchadique) ; la famille ouralienne avec le hongrois (langue finno-ougrienne), la famille altaïque avec le turc (langue turque), la famille sino-tibétaine avec le cantonais (langue chinoise) et enfin deux langues isolées mais dont la classification fait toujours débat : le japonais et le coréen. Les échantillons ont tous été produits par des locuteurs natifs masculins et sont d'une durée de 1,5 à 2 secondes. Ils sont extraits du texte lu « La bise et le soleil » traduit dans les différentes langues. Si le passage en question peut varier d'une langue à une autre, tous ont le même profil intonatif de fin de phrase assertive. Vingt-cinq sujets francophones natifs ont participé aux différents tests. Leur âge était compris entre 14 et 53 ans (moyenne 23 ans). L'expérience se déroulait dans une pièce calme, Les sujets étaient installés devant un écran d'ordinateur et recevaient les stimuli sonores par un casque. Sur l'écran, ils pouvaient voir au centre une grille 16x16 et sur le côté gauche 17 rectangles codés arbitrairement (BB, CC, DD etc.). Pour pouvoir écouter les échantillons, il suffisait aux sujets de double-cliquer sur les rectangles. La consigne qui leur a été donnée était de former des groupes de similarité sonore. Pour ce faire, ils pouvaient écouter les enregistrements autant de fois qu'ils le souhaitaient, et les déplacer à l'aide de la souris sur la grille. Si deux ou plusieurs langues leur semblaient proches du point de vue de leur structure sonore globale, ils devaient les regrouper ensemble sur la grille (en contact), et au contraire disposer les langues dans des groupes différents (sans contact) si celles-ci leur semblaient être suffisamment distinctes. À la suite de ce test, les sujets répondaient à un questionnaire dont le but était d'évaluer leur connaissance pour chacune des langues testées (voir Bra07). Il s'agissait notamment de vérifier que les sujets étaient bien des sujets naïfs, et que leurs réponses dépendaient donc plus de la structure sonore des langues que de leur connaissance des relations génétiques ou géographiques existant entre elles. Les résultats du test de classification ont ensuite été soumis à deux analyses différentes : une analyse en clusters ainsi qu'une analyse MDS (multidimensional scaling).

2.2. Résultats

En ce qui concerne le questionnaire, les réponses pour chaque langue ont été notées de 0 à 2 en fonction leur exactitude : 2 si la lange était identifiée ; 1 si la région géographique ou la famille linguistique était identifiée ; 0 en cas d'absence de réponse ou de réponse erronée. L'arabe a été massivement identifié par 16 sujets sur 25, suivi du japonais (14 sujets) puis de l'hébreu (6 sujets).
En ce qui concerne les zones géographiques ou les familles de langues, les résultats sont assez variables, avec par exemple une excellente reconnaissance de l'appartenance linguistique du cantonais, des langues romanes, germaniques et slaves.
On peut se demander ce que ces résultats nous apprennent sur la signification des réponses obtenues dans les autres tests de l'étude. Par exemple, le taux d'identification élevé de l'arabe pourrait nous inciter à le rejeter de l'échantillonnage des langues. Cependant, nous pensons que les réponses des sujets ne reflètent qu'une représentation vague de l'espace linguistique de l'Afrique du Nord, et que leur identification de l'arabe correspond à une estimation globale sans que la variété dialectale ne soit connue.
L'identification du cantonais comme « chinois » relève sans doute du même genre de jugement approximatif : les sujets n'ont pas identifié une langue ou un dialecte particulier. Ce qui justifie la note 1 (et non 2) pour les réponses « chinois » données pour le cantonais. Le même choix aurait pu être fait pour l'arabe.
En revanche, le japonais correspond bien à une seule et unique variété linguistique, et le taux de reconnaissance très élevé pose effectivement un problème, d'autant plus que le japonais semble être identifié non pas sur des critères phonétiques mais sur des critères lexicaux. En effet, à la fin de l'enregistrement, on peut entendre le suffixe [maʃ ita] qui, s'il n'est pas nécessairement connu du grand public, peut l'être parmi les adeptes de dessins-animés japonais (dont font partie un certain nombre de sujets). En somme, *mashita* en fin d'énoncé est une signature facile à repérer qui "trahit" le japonais.
Une solution conservatrice serait de retirer a posteriori le japonais des langues testées, mais nous avons préféré ne pas agir de la sorte afin de pouvoir comparer nos résultats à ceux obtenus par [Bra07] avec des sujets anglophones sur les mêmes stimuli. Le japonais n'avait en effet pas été écarté dans cette étude (et nous ignorons quel était le taux d'identification du japonais par les anglophones). D'une façon générale, on peut cependant considérer que les connaissances des sujets sur les 17 langues testées, qu'elles soient précises, comme l'identification correcte de telle langue particulière, ou plus vagues, comme l'identification d'une région géographique ou d'une famille linguistique ou les deux, ne sont pas négligeables puisque le score moyen pour le questionnaire était 0.67 sur une échelle de 0 à 2.

Les résultats bruts du test de classification sont une matrice M de dimension 17 x 17 où chaque élément m_{ij} représente le nombre de sujets qui ont classé ensemble les deux langues Li et Lj. La matrice M indique donc, pour chaque couple de langues une valeur de similarité. Nous avons dérivé une matrice de distances D à partir de M : $d_{ij} = (n - m_{ij})/n$ (où n est ici 25, le nombre de sujets). Les analyses en clusters et MDS sont faites sur cette matrice de distances.
L'analyse en cluster (regroupements hiérarchiques) peut produire des résultats différents selon l'algorithme de calcul (ou "méthode") utilisé [Bay08]. Nous présentons dans la Figure 1 une analyse hiérarchique des données de la tâche de classification utilisant la méthode "complete linkage" (la distance entre deux groupes est définie comme la distance maximale entre un élément d'un groupe et un élément de l'autre). Une autre méthode populaire, la méthode de Ward, consiste à calculer un regroupement minimisant la variance intra-groupe [Bay08]. Dans le cas de nos distances perceptives, quelle que soit la méthode utilisée quatre groupes se distinguent nettement : le premier formé de l'hébreu, du néerlandais, du suédois, du slovène et du turc ; le deuxième du haoussa, du croate, du catalan et du galicien ; le troisième du persan, du hongrois, de l'amharique et de l'arabe ; et enfin le quatrième du coréen, du japonais, du cantonais et du sindhi.
L'analyse MDS (multidimensional scaling) [Bor05] qui permet de réajuster sur *n* dimensions les distances entre langues. L'erreur d'ajustement (le STRESS) entre données réelles (la matrice D) et ajustement dépend naturellement du nombre de dimensions *n*. La Figure 2 montre une analyse pour deux dimensions : les langues sont situées dans un espace à deux dimensions et les distances entre langues dans cet espace sont ajustées au mieux avec les distances perceptives observées dans le test de classification. Deux questions se posent immédiatement : (1) Quelle interprétation donner à chacune de ces deux dimensions ? (2) Où placer la langue des sujets, le français, dans cet espace ? Alors que l'analyse MDS permet de représenter les distances entre les langues en deux dimensions, le deuxième test est censé déterminer laquelle de celles-ci est la plus pertinente (dimension 1 ou dimension 2).

3. EXPERIENCE 2

3.1. Méthode

La deuxième tâche, qui lors de l'expérience correspond chronologiquement au troisième test, consiste en une classification des langues en fonction de leur similarité par rapport au français. Le sujet est cette fois face à une échelle sur laquelle il doit placer les 17 langues en fonction de la relation sonore de proximité qu'elles entretiennent avec le français. Il est précisé au sujet qu'il peut choisir de placer plusieurs langues sur la même ligne si celles-ci lui semblent être également distantes du français.

3.2. Résultats

Pour calculer les distances entre les langues, nous avons ordonné les résultats sur une échelle de 1 à 17 pour chacune des configurations réalisée par les 25 sujets. Nous avons ensuite calculé la moyenne et l'écart-type pour chaque langue (Table 1). Ces résultats permettent ainsi d'établir une échelle perceptive qui mesure la distance existant entre le français et les autres langues. Ainsi, le catalan et le galicien semblent perceptivement les plus proches du français, alors que le cantonais est perçu comme étant le plus éloigné. Outre le fait de nous donner des informations intéressantes sur l'agencement perceptif en fonction des langues et des familles linguistiques, les résultats de la table 1 nous permettent de compléter les résultats obtenus au test 1. En effet, en calculant tour à tour le coefficient de corrélation entre les résultats du test 2 avec la dimension 1 puis avec la dimension 2 du test 1, on observe que la dimension 2 corrèle mieux (r =0,68) que la dimension 1 (r = 0,34 ; coefficient de Pearson). Si l'on se penche sur la répartition des langues selon cette dimension (Figure 2), on constate que le critère géographique Nord/Sud permet d'expliquer au mieux la classification perceptive des langues. D'autre part, l'intérêt de confronter les résultats du test 2 au test 1 est de pouvoir placer le français sur l'espace perceptuel de similarité (Figure 3).

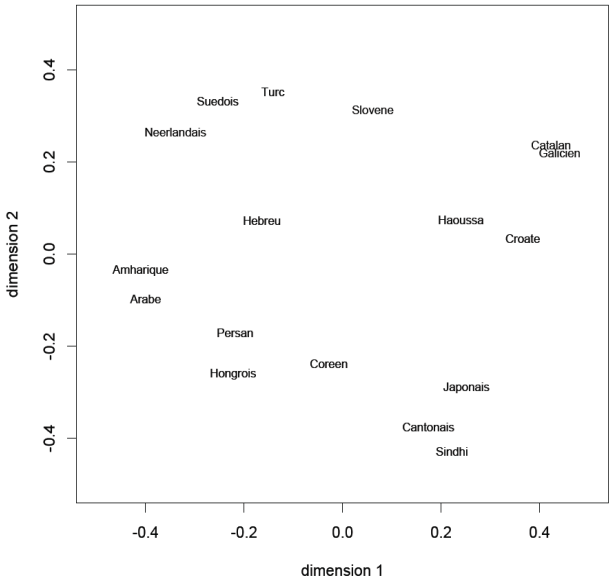

Figure 2. Analyse MDS

4. FIGURES ET TABLES

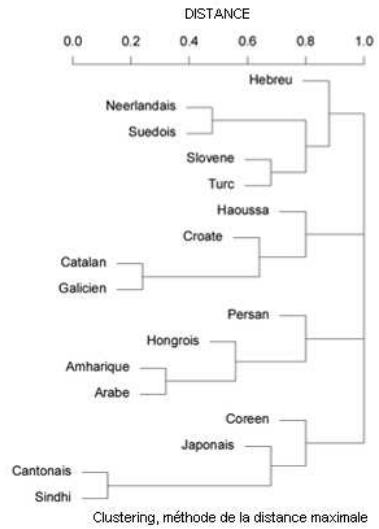

Figure 1. Analyse en clusters pour la tâche de classification

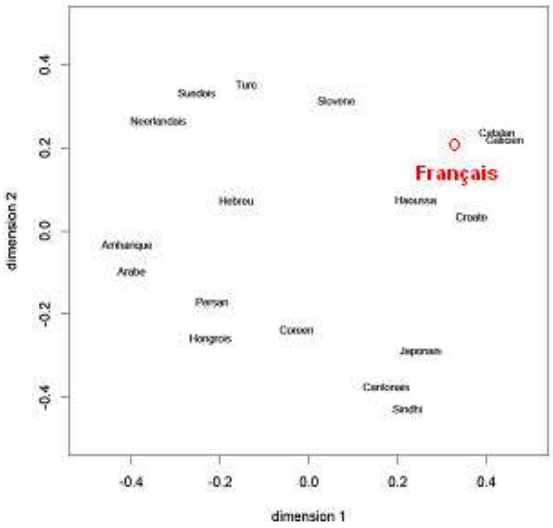

Figure 3. Positionnement du français d'après le test 2

Table 1 : Distance par rapport au français

	Distance moyenne	Ecart-type
Catalan	2,36	3,5
Galicien	2,56	3,6
Croate	5,68	2,1
Suédois	6,12	2,6
Néerlandais	6,16	3,8
Slovène	6,84	3,1
Haoussa	7	2,2
Turc	7,44	4,0
Hongrois	7,56	3,8
Coréen	8,2	3,8
Persan	8,76	3,9
Hébreu	10,48	2,5
Amharique	10,64	3,8
Japonais	11,56	3,9
Arabe	12,48	3,5
Sindhi	12,6	3,2
Cantonais	14,28	3,3

5. CONCLUSION ET DISCUSSION

Tout d'abord, nous avons pensé que l'hypothèse génétique pouvait expliquer en partie cette catégorisation des langues du monde. En effet, si deux langues A et B sont issues d'une même langue commune, elles partagent a priori un certain nombre de caractéristiques communes, qui vont de leurs propriétés syntaxiques, morphologiques, lexicales à leurs caractéristiques phonétiques. De nombreuses lois phonétiques illustrent d'ailleurs ce dernier point (loi de Grimm, loi de Verner, loi de Grassman etc.) et tendent à montrer que la structure sonore des langues évolue, certes graduellement, mais dans une sorte de mouvement commun. Si tous les segments de chacune des langues concernées peuvent ne pas être affectés par de tels processus, il semble donc raisonnable de penser que des langues phylogénétiquement proches doivent partager des traits phonétiques communs, quels que soient leurs caractéristiques propres ou l'ancienneté de leur « lignage ». Qu'en est-il donc de nos résultats ? Le test 1 montre que ceci est vrai pour certaines paires de langues. En effet, le galicien et le catalan ont été majoritairement classés ensemble, ainsi que le suédois et le néerlandais, puis l'arabe et l'amharique, ces trois paires faisant respectivement partie des familles romane, germanique et sémitique. En revanche, de nombreux doutes subsistent pour un certain nombre d'autres paires, comme notamment le sindhi et le cantonais (famille indo-européenne et sino-tibétaine) ainsi que le slovène et le turc (indo-européenne et altaïque) qui ont été classés ensemble. Pour les autres langues, même s'il existe une assez forte variabilité des réponses données par les sujets, on peut se pencher sur les langues avec lesquelles elles ont le plus été regroupées. L'hébreu a par exemple été classé 11 fois avec l'amharique et 13 fois avec l'arabe, ces langues faisant également partie de la branche sémitique

Si l'on se penche maintenant sur le test 2, on devrait voir logiquement apparaître les langues indo-européennes comme les plus proches du français, et ensuite, quels que soient l'ordre des autres familles, une homogénéité à l'intérieur de celles-ci (par exemple, si les langues de la branche sémitique sont jugées comme les plus éloignées du français, on devrait retrouver une certaine homogénéité dans leur répartition étant donné qu'elles font partie d'un même groupe génétique). De nouveau, les résultats sont assez mitigés. Si les langues les plus proches du français sont bien des langues indo-européennes, les deux premières étant de surcroit des langues romanes, le persan et le sindhi apparaissent assez loin dans le classement. De plus, le haoussa, langue tchadique de la famille afro-asiatique, apparaît juste après les langues romanes, germaniques et slaves, alors que l'arabe, l'amharique et l'hébreu sont placés bien plus loin. Bref, au vu de toutes ces remarques, il semble assez difficile de pouvoir valider complètement l'hypothèse de la « filiation » comme la principale motivation des distances perçues entre les langues, même si certaines tendances peuvent être observées, notamment en ce qui concerne l'extrême proximité de la langue de référence avec les langues de son même groupe phylogénétique.

RÉFÉRENCES

[Bra07] Bradlow A. (2007), *A perceptual similarity space for languages*, Proceedings of the XVIth International Congress of Phonetic Sciences, Saarbrucken, Germany.

[Bor05] Borg, I. and Groenen,P. (2005), *Modern Multidimensional Scaling: theory and applications* (2nd ed.), Springer-Verlag New York

[Bay08] Bayen, R.H. (2008), *Analysing Linguistics Data: A practical Introduction to Statistics using R*, Cambridge University Press

[Col85] Collinge, N.E. (1985) *The Laws of lndo-European*, John Benjamins Publishing Co, Amsterdam