Détermination expérimentale d'indices linguistiques pour la discrimination des langues romanes

Ioana VASILESCU, Jean-Marie HOMBERT, François PELLEGRINO

Dynamique du Langage - I.S.H.

14, avenue Berthelot
69363 Lyon Cedex 07 – France

ABSTRACT

This paper deals with perceptual identification and differentiation of five Romance languages, namely French, Italian, Spanish, Portuguese and Romanian. Previous studies have investigated human capability to identify spoken samples in unknown languages after a relatively brief exposure. Accordingly, we conduct an analysis to determine which perceptual categories are salient in Latin languages identification. Three groups of listeners with different mother languages (French, Romanian and Japanese) have been considered. Results reveal that: - identification scores are a function of previous exposure to the languages, - the patterns used for the discrimination within the Latin family are mother tongue dependent and - the segmental cues emerging from the subjects' responses may be relevant in automatic language identification.

1. INTRODUCTION

L'identification automatique des langues a dernièrement bénéficié d'une approche complémentaire issue de l'étude de la discrimination linguistique par des êtres humains. Les résultats témoignent d'une capacité remarquable des auditeurs naïfs à identifier et discriminer des langues complètement inconnues après une période d'apprentissage réduite. Ainsi, les expériences de [Mut94a] ont mis en évidence le fait que même sans exposition préalable à une langue étrangère, les sujets humains sont capables de discrimination. Des taux de 80 % de réussite après une courte écoute (2s) sont rapportés dans [Sto97].

L'information portée par le signal sonore est complexe, et la perception humaine semble capable d'en extraire des traits spécifiques d'une langue à l'autre, en termes d'inventaire phonémique, de règles phonotactiques, de structure intonative ou syllabique. L'étude de la perception des langues par les êtres humains permet non seulement l'amélioration de la compréhension des processus linguistiques de phonologisation, mais fournit également de nouveaux traits pertinents dans le cadre des systèmes automatiques d'identification des langues. Les systèmes de reconnaissance actuels sont essentiellement basés sur la modélisation phonotactique et il semble évident que de nouveaux traits soient nécessaires afin d'améliorer les taux de réussite.

Plusieurs facteurs ont été analysés au travers des expériences en identification des langues par les sujets humains. Ils concernent le matériel linguistique, la structuration de la tâche ou les caractéristiques des sujets en termes de connaissances linguistiques préalables. Le facteur "matériel linguistique" a permis la mise en évidence de stratégies perceptuelles face un nombre important (10) de langues de test [Mut94b]. La variation des tâches (discrimination vs. évaluation de la proximité des langues) a permis l'émergence de traits linguistiques discriminants [Sto96]. Le facteur "population" a révélé des différences dans la discrimination des langues en fonction du temps d'apprentissage ou du caractère monolingue ou bilingue des sujets [Mar99].

Notre étude se propose de vérifier dans quelle mesure la langue maternelle des sujets participants au test influence le type d'information linguistique que ces derniers vont utiliser pour la discrimination de 5 langues romanes, à savoir le français, l'italien, l'espagnol, le roumain et le portugais. Trois populations ont participé à l'expérience : deux d'entre elles étaient constituées de locuteurs natifs d'une langue romane (Français et Roumains) et la troisième, ayant le rôle de population de contrôle était constituée de sujets Japonais ne possédant aucune exposition préalable aux langues romanes.

Nous allons par la suite procéder à une brève présentation de la famille des langues romanes et des traits segmentaux et supra-segmentaux *a priori* pertinents pour leur discrimination. Dans le Section 3 nous présentons le corpus linguistique, la tâche de discrimination et les caractéristiques des populations participantes. Finalement, la 4ème Section décrit les résultats et fournit leur interprétation.

2. LES LANGUES ROMANES

Les langues romanes sont les idiomes issus du latin, un dialecte italique faisant partie de la branche occidentale de la famille indo-européenne, à savoir la branche italo-celtique [Ruh91]. La famille romane réunit 5 des langues les plus utilisées dans le monde contemporain. Elles représentent la langue ou l'une des langues officielles dans 7 pays européens (Belgique, France, Espagne, Portugal, Suisse, Italie et Roumanie), un continent (Amérique du Sud) et d'autres régions du monde (Canada, Amérique centrale etc.).

Notre approche prend en considération 5 langues romanes : le français, l'italien, l'espagnol, le port ugais et le roumain. Ces langues sont représentatives de la famille romane du point de vue économico-politique, mais aussi de la distribution géographique (langues romanes occidentales : le français, l'italien, l'espagnol et le portugais vs. orientales : le roumain) qui entraîne des particularités spécifiques dues à des interactions avec d'autres langues (germaniques, slaves etc.). Une description phonologique de ces langues révèle en même temps que des traits communs génétiquement explicables, de nombreuses particularités qui concernent à la fois le niveau segmental et supra segmental.

La structure des systèmes vocaliques partage les langues romanes en deux groupes : celui des langues possédant deux oppositions vocaliques, antérieur/postérieur (italien, espagnol) et celui des langues possédant trois oppositions vocaliques, antérieur/postérieur/central ou antérieur-arrondi (roumain, français, portugais). De plus, la famille romane inclue deux des quatre langues européennes possédant des voyelles nasales phonologiques, moyennes et ouvertes (les deux autres langues étant le polonais et certains dialectes bretons) [Ruh74]. Les systèmes consonantiques sont plus homogènes en termes de traits communs, néanmoins des segments spécifiques témoignent des évolutions individuelles (par exemple, la consonne fricative glottale /h/ en roumain, la fricative dentale /θ/ en espagnol, etc.).

L'analyse du niveau supra segmental partage les langues romanes en 4 groupes en fonction de leur structure rythmique: syllabiques (l'italie n et le roumain), accentuelle (le portugais), "trailer-timed" (l'espagnol) et à accent fixe ou "langue de frontière" (le français) [Hir98].

3. MATERIEL ET METHODE

3.1. Corpus linguistique

Les enregistrements (22kHz, chambre insonorisée, intensité normalisée) de 4 locuteurs, 2 hommes et 2 femmes, pour chacune des langues ont été utilisés pour cette expérience. Deux d'entre eux (homme et femme) ont été utilisés dans la phase d'apprentissage, les deux autres ayant servi à la constitution des stimuli de test. Le corpus inclut de la parole lue et des histoires quasi spontanées.

3.2. Populations

20 Français, 20 Roumains et 20 Japonais, hommes et femmes, âgés de 18 à 60 ans et ayant au moins un niveau de formation correspondant au baccalauréat, ont participé au test. L'exposition préalable aux langues de test est homogène pour chacun des groupes :

Les Français ont étudié l'espagnol à l'école, mais aucun d'entre eux ne parlait couramment cette langue ou aucune des autres langues romanes. De plus, la France est géographiquement en contact avec l'Espagne et l'Italie.

Les Roumains ont étudié le français à l'école, mais aucun d'entre eux ne le parlait couramment ou aucune autre langue romane. La Roumanie n'est voisine d'aucun des pays de langues romanes. Cependant des fictions télévisées produites en Amérique latine sont souvent diffusées sur les chaînes nationales en version originale sous-titrée.

Les Japonais n'ont étudié aucune des langues romanes et aucune exposition préalable à ces langues n'a été mentionnée.

3.3. Conditions d'expérimentation

L'expérience a été divisée en trois phases :

- L'entraînement a permis aux sujets de se familiariser avec chacune des langues romanes. Il a consisté dans l'écoute de deux extraits de 10s dans chaque langue. Les extraits, prononcés par deux locuteurs différents, un homme et une femme, ont été présentés en ordre aléatoire.
- Durant le test proprement dit, les sujets devaient prendre une décision de type "même langue"/ "langue différente" pour chaque item. 50 stimuli de type AB ont été présentés : Chaque stimulus durait en moyenne 6s et était séparé du second par un court son de type « cloche ». Le sujets disposaient de 2s après chaque séquence AB pour répondre si A et B provenaient de la même langue ou de langues différentes. Les extraits étaient présentés une seule fois et chaque combinaison L_i-L_j, où $\{i,j\} \in [1,\ldots,5]^2$ a été présentée deux fois.
- A la fin du test les sujets ont eu la possibilité de s'exprimer sur la nature des indices qui les ont aidé à discriminer les langues.

4. RESULTATS

4.1. Analyse préliminaire de la significativité

Un test de significativité statistique des réponses des trois populations a été effectué avant de procéder à leur analyse multidimensionelle. L'histogramme présenté dans la Figure 1 reproduit le pourcentage de réponses correctes pour chaque paire de langues, ainsi que les paires pour lesquelles les réponses ont été significatives ou non (t-test univarié, $p<0,001$). Les réponses des populations française et roumaine ont été significatives dans la plupart des cas (sauf les paires Portugais/Portugais pour les deux populations et Roumain/Portugais et Portugais/Roumain pour la population française). Les réponses des Japonais ont été données dans la majorité des cas au hasard. Par conséquent une analyse multidimensionelle de leurs réponses n'est pas pertinente ; elle n'est donc pas présentée dans la suite.

4.2. Analyse multidimensionnelle

Les analyses multidimensionnelles ont été effectuées à l'aide du logiciel Vista [Vis99] pour les réponses des populations française et roumaine.

Les sujets français. Les réponses ont reçu une représentation tridimensionelle. La Figure 2 montre la projection des résultats dans deux plans définis, le premier (D1/D2) par les deux dimensions principales, et le seconde par la première et la troisième dimension (D1/D3). Dans le plan D1/D2 trois groupes de langues apparaissent : la langue maternelle, les langues familières et les langues inconnues. La première dimension sépare la langue maternelle (français) des autres idiomes, tandis que la seconde permet de distinguer entre langues familières (italien, espagnol) et inconnues (portugais, roumain). La troisième dimension distingue plus nettement entre les langues familières, tandis qu'entre les langues inconnues la confusion se maintient. Il semble, par conséquent, que les sujets français soient difficilement capables de distinguer entre deux langues inconnues après une période très courte d'apprentissage. Finalement, des considérations phonologiques devraient aussi être prises en considération, dans la mesure où la seconde dimension semble séparer les langues en fonction de la complexité de leurs systèmes vocaliques : les langues à trois oppositions vocaliques (portugais, roumain) sont séparées des langues à deux oppositions (espagnol, italien).

Les sujets roumains. L'analyse des réponses à reçu également une représentation tridimensionnelle (Figure 3). La première dimension distingue la langue maternelle des autres langues romanes. La seconde dimension sépare les langues articulées autour de deux oppositions (italien, espagnol) des langues dont le système s'organise autour de trois oppositions (français, portugais). Le plan D1/D3 semble correspondre à une distribution géographique des langues, isolant les langues ibériques (espagnol, portugais) des autres langues romanes. Elle pourrait être également la conséquence de l'exposition fréquente à l'espagnol et portugais sud-américains au travers les fictions télévisées. La troisième dimension séparerait les langues familières (espagnol et au portugais) des langues moins familières (français et surtout italien qui n'est pas appris à l'école).

Les sujets japonais. Ils ont répondu au hasard, confirmant ainsi qu'une exposition préalable aux langues facilite grandement la tâche de discrimination. Des expériences impliquant un apprentissage plus long sont envisagées .

5. PERSPECTIVES

La présente étude visait à étudier les traits discriminants entre 5 langues romanes au travers d'une expérience perceptuelle. Elle met en évidence des stratégies différentes de discrimination des langues inconnues. La perception des langues étrangères semble être filtrée par les traits de la langue maternelle et repose sur différents types d'informations, linguistique (segmentale et supra segmentale) et/ou extra linguistique (notamment socio-linguistique pour ce qui est des sujets roumains qui doivent leur connaissances sur les langues ibériques aux médias). De plus, une exposition antérieure aux langues romanes et la connaissance d'au moins une de ces langues sont des facteurs fondamentaux dans leur discrimination. Les sujets réellement naïfs (les Japonais) sont incapables d'extraire des paramètres saillants après une courte exposition à la langue (20s).

Nos prochaines démarches tenterons de mieux définir les niveaux linguistiques responsables de la discrimination, de donner une meilleure définition à la notion d'"exposition à la langue" et, finalement, de valider les indices mis en évidence par l'approche expérimentale dans un système de reconnaissance automatique.

6. REMERCIEMENTS

Les auteurs remercient Sumikazu Nishio pour son aide dans la mise en place du protocole expérimental auprès des sujets japonais.

BIBLIOGRAPHIE

[Hir98] Hirst D., DiCristo A. (1998), Intonation Syste. A Survey of Twenty Languages, Cambridge University Press.

[Mar99] Marks E.A., Bond Z.S., Stockmal V. (1999) "The effect of proficiency in a specific foreign language on the ability to identifiy a novel foreign language", 14 th ICPhS, pp. 133-135.

[Mut94a] Muthusamy I.K., Barnard E., Cole R.A. (1994), " Automatic language identification: A review/Tutorial", IEEE Signal Processing magazine.

[Mut94b] Muthusamy I.K., Jain N., Cole R.A. (1994), " Perceptual benchmarks for automatic language identification ", IEEE ICASSP.

[Ruh74] Ruhlen M. (1995), " Some comments on vowel nasalization in French. Notes and discussion ", Journal of Linguistics,10.

[Ruh91] Ruhlen M. (1995), Guide to the World's Languages, Stanford University Press.

[Sto94] Stockmal V., Muljani D., Bond Z. (1994), " Can children identify samples of foreign languages as same or different?", Language Sciences, 16, pp. 237-251.

[Sto96] Stockmal V., Muljani D., Bond Z. (1996), " Perceptual Features of Unknown Foreign Languages as Revealed by Multidimensional Scaling", ICSLP.

[Vis99] The Visual Statistics System Vista web page http://forrest.psych.unc.edu/research/ (visitée en Novembre 1999)

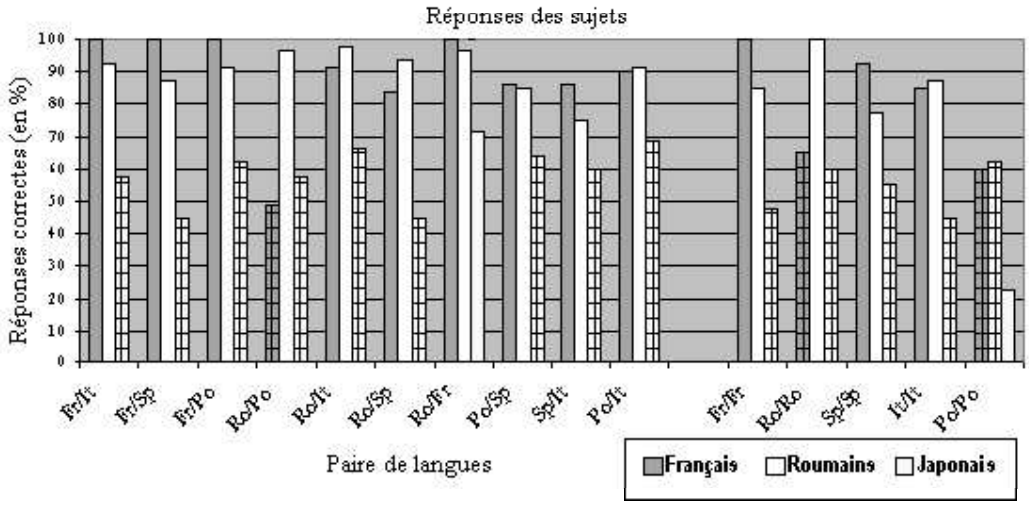

Figure 1 – Réponses correctes pour chaque groupe de sujets (Français, Roumains et Japonais). Les abscisses indiquent les paires de langue AB (AB et BA sont cumulés). Une barre pleine (resp. rayée) indique un score significatif (resp. non significatif) avec $p < 0{,}001$.

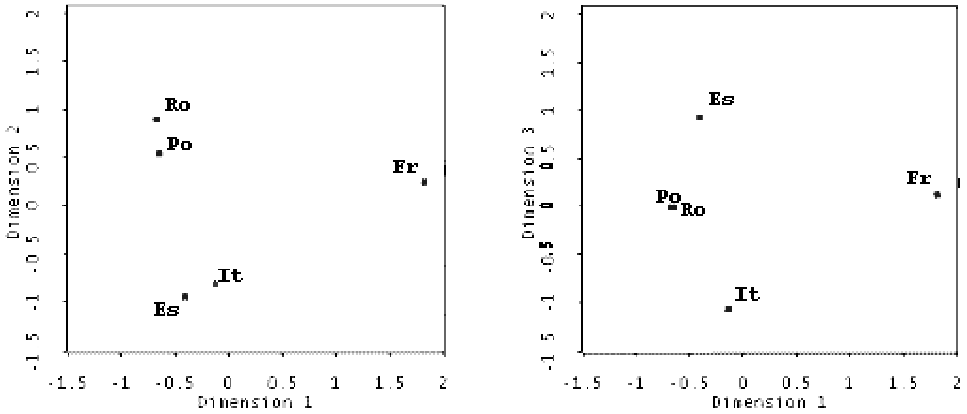

Figure 2 – Projection des réponses des sujets **Français** dans un espace multidimensionnel. A gauche la projection est réalisée dans le plan des deux premières dimensions et à droite dans le plan Dimension 1 / Dimension 3.

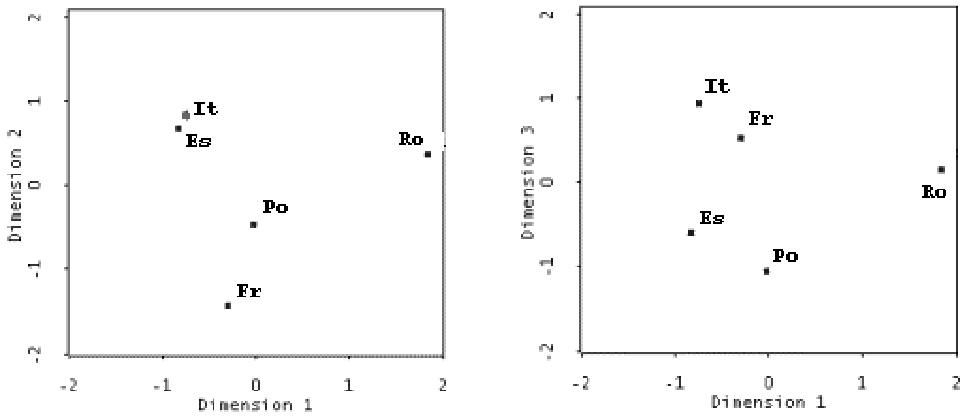

Figure 3 – – Projection des réponses des sujets **Roumains** dans un espace multidimensionnel. A gauche la projection est réalisée dans le plan des deux premières dimensions et à droite dans le plan Dimension 1 / Dimension 3.