Stratégies perceptives en identification des langues

Ioana VASILESCU

LIMSI-CNRS
BP 133 Bât. 508 – 91 403 Orsay Cedex, France
Tél.: ++33 (0)1 69 85 81 25 - Fax: ++33 (0)1 69 85 80 88
Mél: Ioana.Vasilescu@limsi.fr - http://www.limsi.fr/

ABSTRACT

This paper deals with perceptual strategies in language identification. The study of strategies employed by humans to identify foreign languages is currently considered as a comparative approach in evaluating automatic performances. We present a survey of the domain and suggest a methodology aiming to control the factors responsible of the identification scores, *i.e.* experimental design, corpus and listeners' linguistic background. Two experimental designs are conducted (language discrimination *vs.* evaluation of the similarity) to determine the strategies developed by four populations to identify Romance languages. We will present here only a case study (American subjects) highlighting the main identification strategies (linguistic cues *vs.* previous exposure to the languages).

1. INTRODUCTION

Durant la dernière décade, nous avons été témoins d'un intérêt croissant pour la comparaison Homme/Machine en identification des langues [Lip97], [Pol99]. Cet intérêt est justifié par l'objectif de modéliser les stratégies perceptives afin d'améliorer les performances actuelles des systèmes automatiques. Néanmoins, malgré la prolifération des études consacrées à l'identification des langues par les humains, nous ne pouvons pas parler d'un gain réel en termes de réussite automatique due à la valorisation de ces stratégies. En effet, les résultats des démarches expérimentales en identification sont à ce jour quasi inexploitables. Une explication potentielle à cet état des choses pourrait se trouver dans l'hétérogénéité des méthodes employées, qui ne permettent pas pour l'heure, d'une part, la convergence des résultats vers une typologie des stratégies perceptives utilisées par les humains et, d'autre part, une meilleure prise en compte des stratégies proprement linguistiques. Nous estimons que l'hétérogénéité se trouve à la fois dans la diversité des paradigmes expérimentaux employés, dans le choix des langues constituant les corpus de test et dans le profil linguistique des auditeurs.

Quatre *paradigmes* sont couramment employés : identification, discrimination, évaluation de la similarité des langues et recherche d'intrus. Le paradigme d'identification est le plus usité, car étant considéré comme le plus proche aussi bien des tâches automatiques que des circonstances réelles d'écoute de langues [Mut94]. Il consiste en la présentation d'une suite de stimuli sonores, généralement en parole naturelle, que les auditeurs doivent identifier en termes de langue d'origine. La durée des stimuli est variable, allant d'une seconde à une dizaine de secondes, et le test est souvent précédé d'un bref entraînement permettant aux sujets de se familiariser avec les langues. Par la suite, on évalue les pourcentages d'identification correcte. Le paradigme de discrimination consiste en la présentation d'échantillons sonores appariés. On demande aux sujets de se prononcer sur l'origine des stimuli en précisant s'il s'agit d'échantillons de parole issus d'une même langue ou de deux langues [Bon91], [Sto96]. Quant au paradigme d'évaluation de la similarité, les stimuli sont présentés de la même façon que lors du paradigme de discrimination, mais la requête est différente : il s'agit d'évaluer la ressemblance des échantillons de parole au moyen d'une échelle de valeur [Sto96], [Sto99]. Enfin, le paradigme de recherche d'intrus consiste en la présentation de stimuli sous la forme d'une suite AAX, où AA représente le contexte et X peut être A ou B [Ram99]. Il est particulièrement adapté aux stimuli en parole modifiée, car exigeant des sujets une comparaison immédiate des extraits et non pas une association stimulus/langue, plus coûteuse en termes de mémoire et d'attention.

Les *langues* généralement choisies sont un groupement hétéroclite qui privilégie les langues européennes. La constitution du corpus est motivée davantage par la diversité que par les particularités structurelles des langues. Un exemple illustratif est le corpus OGI, qui contient de la parole téléphonique spontanée en 22 langues et se trouve parmi les premiers à fournir des stimuli pour des tests perceptifs en identification [Mut94].

Concernant le profil linguistique des *auditeurs*, la plupart des études s'intéressent exclusivement aux anglophones ; d'où la difficulté de généralisation des stratégies perceptives observées. Notons également la variabilité en termes d'exposition antérieure des sujets à d'autres langues, généralement non vérifiée et qui entraîne des scores de reconnaissance difficilement interprétables.

Finalement, des *indices* discriminants ont été mis en évidence. Pour l'heure, les indices prosodiques ont

bénéficié d'un intérêt plus particulier et se sont avérés spécialement robustes [Ram99] [Bar99].

Concluons donc que la comparaison des résultats et la mise en évidence d'une typologie des stratégies perceptives et d'indices linguistiques discriminants représentent une tâche complexe. La généralisation des informations obtenues grâce aux différents paradigmes cités est complexifiée par le choix arbitraire (anglophone) des populations de test et par la diversité des langues composant les corpus.

Pour notre part, nous nous sommes intéressée à la valorisation des acquis mentionnés en identification perceptive des langues afin de rendre les résultats exploitables. Il s'agissait, d'une part, de contrôler les facteurs jouant dans l'obtention des résultats, et d'autre part, de faire converger les observations issues de configurations expérimentales différentes vers une typologie des stratégies perceptives. Le bilan suggère qu'au moins trois facteurs influencent les résultats: le type de paradigme expérimental employé, le profil linguistique des populations et les particularités structurelles des langues. Ci-dessous nous présentons la méthodologie développée (2) et une étude de cas (3).

2. DISCRIMINATION *VS.* ÉVALUATION DE LA SIMILARITÉ DES LANGUES

L'objectif de notre démarche a été de mettre en évidence des stratégies perceptives en identification des langues. Nous avons tout d'abord opéré un choix délibéré et homogène des idiomes de test en faisant appel à un critère génétique de sélection des langues, *i.e.* nous avons inclus dans l'expérience cinq langues de la famille romane : français, italien, espagnol, portugais et roumain. L'idée sous-jacente à ce choix est celle que la communauté de traits due à l'origine commune des langues devrait complexifier la tâche expérimentale. Nous avons fait appel à deux types de population de test, d'après le critère "langue maternelle" des sujets (*i.e.* [+/- langue maternelle romane]). Ce choix a été fait afin de pouvoir distinguer entre des stratégies d'identification dépendantes de la langue maternelle et/ou des langues familières des sujets, et des stratégies linguistiques, dépendantes des indices à fort caractère discriminant, perçus dans les stimuli.

Nous avons opté pour deux paradigmes expérimentaux : discrimination et évaluation de la similarité sonore des langues. Le paradigme de discrimination est adapté d'après celui de recherche d'intrus ; on demande aux auditeurs de différencier et non pas d'identifier les langues. Le but est d'observer lesquels des idiomes sont les plus confondus, de déterminer pourquoi certaines langues sont plus associées que d'autres, et, en dernier lieu, de délimiter les critères linguistiques responsables de la discrimination sans difficulté des autres idiomes. De plus, la parenté génétique des langues, impliquant le partage d'un nombre de traits communs, nous amène à prédire une pertinence singulière des indices discriminants découverts. Dans un second temps, le paradigme d'évaluation de la similarité des langues permet de confirmer les jugements *implicites* sur leur ressemblance (source de confusions inter linguistiques), observables par le biais du paradigme de discrimination. En effet, ce paradigme laisse aux auditeurs l'opportunité de "faire des confusions" (*i.e.* de fournir des jugements *explicites*) volontairement et de valoriser ainsi les connaissances acquises sur les langues romanes lors de l'apprentissage proposé ou avant l'expérience. À terme, les indices linguistiques différenciateurs repérés par le paradigme de discrimination peuvent être considérés comme efficients seulement dans la mesure où le paradigme d'évaluation de la similarité confirme les mêmes ressemblances perceptives.

2.1 Corpus

Une base de données acoustiques a été élaborée par l'enregistrement de quatre locuteurs par langue (hommes et femmes). Trois supports ont été utilisés : un livre d'images dont les locuteurs ont décrit la trame narrative [May69] ; la base de données EuRom4 [Eur97] consistant dans la lecture de journaux ; une histoire personnelle du locuteur. Les enregistrements ont été digitalisés à 22kHz, 16 bits, monophonique, sous *SoundForge©*. 10 échantillons de 10s ont été extraits pour la phase d'entraînement, et 100 échantillons (soit 2 [locuteurs] x 5 [par langue] x 5 [langues de test]) de 6s, présentés en paire dans toutes les combinaisons possibles, pour la phase de test.

2.2 Sujets

Deux types de populations ont été choisies selon le critère [+/– langue maternelle romane]. L'expérience de discrimination a été réalisée par des Français, Roumains *vs.* Japonais, Américains. L'expérience d'évaluation de la similarité des langues a été réalisée par des Français *vs.* Américains, autres que ceux de l'expérience précédente. Chaque groupe comportait 20 sujets, testés individuellement dans leurs pays d'origine.

2.3 Protocoles

Les expériences ont été réalisées à travers une interface programmée en PsyScope. Chaque expérience a consisté en deux phases : a. apprentissage, permettant une familiarisation avec les langues de test (10 échantillons de 10s, deux par langue) ; b. test, consistant en l'écoute de 50 paires de signaux. Les échantillons étaient séparés par un bruit ("cloche"), et après chaque paire les sujets disposaient de 2s pour juger : (1) si les deux signaux étaient de type "même langue"/"langues différentes" (discrimination) ; (2) du degré de similarité des deux signaux sur une échelle de

1 à 5 (où 1=pas similaire, 5=très similaire) (évaluation de la similarité). Le délai de 2s expiré, on annonçait aux sujets par un autre bruit ("bip"), qu'une nouvelle paire allait suivre.

3. EXEMPLE : LE CAS DES SUJETS AMÉRICAINS

Quatre différentes populations ont participé aux expériences (*2.2*). Nous allons discuter ici, à titre d'exemple, uniquement le cas des sujets américains (anglophones, monolingues).

3.1 Discrimination des langues romanes

Les résultats en discrimination sont présentés dans les figures 1a et b (% de réussite) et 2a (distances perceptives inter langues). Ils mettent en évidence une primauté des stratégies linguistiques, suivies par celles issues de la familiarité avec certains idiomes du test (français, espagnol). Les premières sont illustrées par les scores obtenus pour les paires de langues Portugais / Français et surtout Espagnol / Italien. Ce dernier item est reconnu en proportion de 67,5%[1] ce qui implique un écart statistiquement significatif par rapport à tout item de test comportant un échantillon en espagnol (reconnus à plus de 80%). Il en est de même pour la paire Français / Portugais, comparée aux autres items comportant du français.

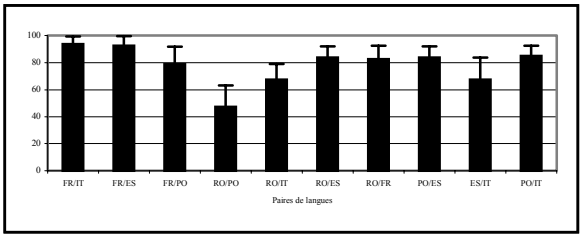

Figure 1a : Scores de discrimination correcte pour les paires de type "langue différente".

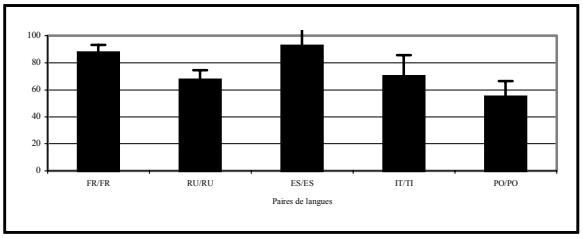

Figure 1b : Scores de discrimination correcte pour les paires de type "même langue".

De ce fait, nous pouvons avancer l'hypothèse que les auditeurs anglophones ont dû repérer des caractéristiques communes des langues citées (italien, espagnol *vs.* français, portugais, roumain) qui ont entraîné des confusions, malgré leur familiarité *a priori* avec l'espagnol et le français. En revanche, la familiarité avec ces deux langues a joué un rôle essentiel dans leur discrimination des autres langues romanes. Ainsi, la Figure 1a montre des scores supérieurs à 80% pour les paires Espagnol / Roumain, Espagnol / Français, Espagnol / Portugais, Français / Roumain et Italien / Français.

Ces données sont plus distinctement mises en valeur par une analyse multidimensionnelle. Ainsi, la première dimension sépare l'espagnol et l'italien, du roumain, du portugais et du français (Figure 2a). La seconde dimension sépare l'espagnol et le français des trois autres langues romanes. Elle peut être associée à une stratégie non linguistique (*i.e.* la familiarité), étant donné que les langues plutôt connues des Américains sont isolées des autres langues romanes. En revanche, la première dimension peut être associée à un indice linguistique et confirme une tendance observée chez la totalité des populations mentionnées dans *2.2*. Elle met en valeur deux groupes linguistiques distinctes : espagnol, italien *vs.* roumain, portugais, français.

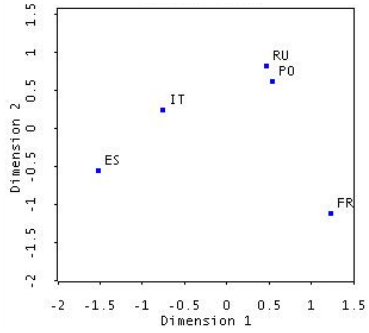

Figure 2a : Distances perceptives selon les deux premières dimensions (D1/D2) (discrimination).

3.2 Evaluation de la similarité des langues romanes

Le paradigme expérimental d'évaluation de la similarité sonore des langues romanes a confirmé la distribution antérieure (Figure 2b).

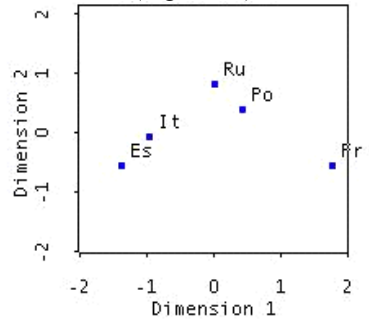

Figure 2b : Distances perceptives selon les deux premières dimensions (D1/D2) (évaluation de la similarité).

Alors que la seconde dimension renvoie à la familiarité, la première départage deux groupes linguistiques : italien, espagnol *vs.* français, portugais ; le roumain peut être à la fois considéré comme neutre

[1] t-test univarié, $p<0.001$.

par rapport à la distribution et appartenant au groupe qui réunit le français et le portugais. Les proximités observées dans *3.1* sont ainsi explicitement validées par un jugement sur les ressemblances sonores des idiomes. Autrement dit, l'indice responsable de la distinction des deux groupes linguistiques peut être estimé comme pertinent, indépendamment de la tâche expérimentale utilisée. Par ailleurs, la synthèse des résultats des autres populations a abouti à une configuration des distances linguistiques similaire à celle fourni par les Américains [Vas00].

A ce stade de notre travail, nous avons considéré les interprétations potentielles quant à la nature de l'indice discriminant, responsable des regroupements obtenus. Etant donnée la complexité de l'information portée par des stimuli en parole naturelle et l'absence d'études consacrées au rôle respectif des différents indices perceptifs d'identification des langues, en général, et des langues romanes, plus particulièrement, nous avons fait appel aux approches typologiques qui étudient les critères linguistiques distinguant les langues romanes. Le bilan des travaux consacrés aux indices segmentaux, supra segmentaux et phonotactiques nous a conduit à formuler l'hypothèse concernant le rôle potentiel du vocalisme pan roman. Le trait linguistique possiblement discriminant serait de type [+/- système vocalique complexe] [Vas00], [Vas01]. Ainsi, le groupe des idiomes possédant des systèmes vocaliques quasi prototypiques (italien, espagnol), serait isolé des langues ayant développé un nombre d'oppositions spécifiques (français, portugais, roumain) [Ter85].

4. CONCLUSIONS

Cette étude a été consacrée au bilan des différentes approches en identification perceptive des langues et à la recherche de stratégies employées par les auditeurs naïfs pour différencier des langues étrangères. Nous avons souligné que plusieurs facteurs jouent dans l'identification perceptive des langues, liés aussi bien aux particularités structurelles des idiomes qu'au profil linguistique des auditeurs. Quant aux stratégies développées par ces derniers, afin d'accomplir la tâche expérimentale, nous avons mentionné le rôle de la familiarité antérieure avec les langues testées et de la pertinence de certains traits linguistiques discriminants. Pour ce qui est des langues romanes, nous avons avancé l'hypothèse du rôle potentiel de la complexité de leurs systèmes vocaliques. Cependant, il s'agit d'une hypothèse qui devra être vérifiée en utilisant, par exemple, des stimuli en parole modifiée où les informations vocaliques exclusivement sont préservées. Par ailleurs, le rôle potentiel des autres indices segmentaux, supra segmentaux ou encore phonotactiques, devra être considéré au travers des constructions expérimentales appropriées.

Ultérieurement, nous nous pencherons sur la comparaison du succès des humains et des systèmes automatiques dans différentes configurations expérimentales. Il sera question de déterminer les configurations où les stratégies perceptives s'avèrent supérieures et d'envisager leur étude dans la perspective d'une modélisation automatique.

BIBLIOGRAPHIE

[Lip97] Lippmann J.L. (1997), "Speech recognition by machines and humans", Speech Communication, Vol. 22, pp.1-15.

[Pol99] Pols L. C. W. (2001) "Flexible, robust, and efficient human speech processing versus present-day speech technology", ICPhS'99.

[Mut94] Muthusamy I.K., Barnard E., Cole R.A. (1994), "Automatic Language Identification: A review/tutorial", IEEE Signal Processing Magazine, Vol. 11, pp. 33-41.

[Bon91] Bond Z., Fokes, J. (1991), "Identifying foreign languages", ICPhS'91.

[Sto96] Stockmal V., Muljani B., Bond Z. (1996), "Can children identify samples of foreign languages", Language Sciences, Vol. 16, pp. 237-254.

[Sto99] Stockmal V., Bond Z. (1999), "Rhythm and region : scaling the perceptual dimensions of Korean ", ICPhS'99.

[Ram99] Ramus, F. (1999), "Rythme des langues et acquisition du langage", Thèse de Doctorat Nouveau Régime en Sciences Cognitives, EHESS.

[Bar99] Barkat, M., Ohala, J., Pellegrino, F. (1999), "Prosody as a distinctive feature for the discrimination of Arabic dialects", ICPhS'99.

[Ter85] Ternes E. (1985), "Typologie des langues romanes du point de vue phonétique et phonologique", Actes du XVIIème Congrès International de la Linguistique et Philologie Romanes.

[May99] Mayer M. (1969) Frogg, where are you ? Dial Books for Young readers, New-York.

[Eur97] EuRom4 (1997), Manuel de grammaire constrastive des langues romanes, Blanche-Benveniste C. ed.

[Vas00] Vasilescu I., Pellegrino F., Hombert J.M. (2000), "Perceptual features for the identification of Romance languages", ICSLP'00.

[Vas01] Vasilescu I. (2001), "Contribution à l'identification automatique des langues romanes", Thèse de Doctorat Nouveau Régime en Sciences du Langage, Université Lumière Lyon 2.