Vers une mesure de la complexité des systèmes phonologiques

Egidio Marsico, Christophe Coupé, François Pellegrino, Jean-Marie Hombert
Dynamique Du Langage, Lyon

1. Introduction

Les descriptions typologiques des systèmes phonologiques permettent de définir et hiérarchiser des classes homogènes de systèmes en fonction de leur contenu. Même si ces classes peuvent être définies de manière qualitative -quels segments- ou quantitative -combien de segments- rien ne permet actuellement de comparer des systèmes de contenus différents en terme de complexité structurelle.

Nous proposons d'évaluer l'architecture des systèmes phonologiques au travers de mesures de complexité et de distance.

La principale difficulté relative aux notions de distance et de complexité, est l'identification d'indices de mesure exogènes, qui ne sont pas des transformations explicites des intuitions des linguistes, mais bel et bien des éléments externes aux systèmes et dont les contraintes seront finalement transposables à l'état observé des systèmes phonologiques. Si un phonologue est prêt à affirmer qu'un système vocalique à 10 voyelles (/i, e, a, o, u/ plus les contreparties nasales) est moins complexe qu'un système à 10 timbres différents, nous ne pouvons introduire ce fait en tant que mesure 'directe' de la complexité, nous devons produire une mesure à même de confirmer cette intuition sans l'inclure en primitive. Avec le même souci d'objectivité, nous ne pouvons considérer la distribution fréquentielle d'un type comme un indice de sa complexité, et partant, comme une pondération dans les calculs de distance, puisque nous aboutirions alors à des définitions circulaires du type "c'est moins complexe car plus fréquent". D'un point de vue théorique, cela consisterait à introduire dans la définition des mesures, les propriétés à faire émerger.

Puisque nous ne pouvons réduire notre mesure de distance à un simple différentiel du nombre de segments, il faut alors prendre en compte la complexité interne des systèmes (nature des segments impliqués). Dès lors, il faut déterminer une pondération objective des segments afin de différencier les systèmes à nombre égal d'éléments. Nous avons traduit cette recherche de hiérarchisation des segments par une recherche de hiérarchie des traits.

La question est alors d'identifier les poids respectifs des différentes dimensions articulatoires retenues, est-ce qu'une implosive est plus complexe qu'une occlusive ? Est-ce qu'une affriquée est plus simple qu'une occlusive aspirée ?

2. Elaboration de la mesure de distance

Deux cadres théoriques semblaient à même de pourvoir une telle hiérarchie, tout d'abord la géométrie des traits (Clements, 1985) qui, avec une définition arborescente des segments permet d'accéder à une pondération objective des dimensions, où le calcul de distance peut être envisagé par la différence des arbres.

La phonologie des gestes articulatoires (Browman & Goldstein, 1989) offre quant à elle, des primitives ancrées directement dans les contraintes du tractus vocal, plus que ne l'est a priori un trait comme 'affriquée' qui sous un seul terme encode un ensemble de gestes.
Néanmoins, nous avons dû renoncer à ces cadres théoriques pour deux raisons, premièrement ils ne couvraient pas l'ensemble des traits que nous utilisons dans la description des segments et deuxièmement, ils ne permettaient pas d'aboutir à une seule et unique hiérarchie. Nous nous sommes donc orientés vers une approche à base de réseaux de neurones.

Nous avons donc élaboré un réseau simple dans lequel chaque neurone correspond à un trait distinctif différent. Au départ tous les neurones sont reliés entre eux et les synapses ont un poids nul. Afin d'initialiser le réseau et de pondérer les synapses (autrement dit de hiérarchiser les traits), nous donnons en entrée au réseau l'ensemble des 930 segments utilisés dans la description des 451 systèmes phonologiques d'UPSID. Il s'agit bien ici d'utiliser l'ensemble des segments possibles et non pas leur fréquence d'apparition dans les langues. Le principe est alors simple, puisque chaque segment est défini par un ensemble de traits, chaque fois qu'un segment est fourni au réseau les synapses reliant les traits utilisés pour le décrire sont incrémentées de 1, et ainsi de suite pour les 930 segments. A la fin de cette phase d'initialisation, toutes les synapses n'ayant pas été activées sont supprimées et les autres reçoivent un poids Ni,j égal au quotient de leur nombre d'activation par 930. Le réseau résultant est alors employé dans le calcul de distance.

Pour calculer la distance entre deux langues L1 et L2, nous procédons comme suit :

1) nous passons dans le réseau initialisé l'ensemble des segments de la langue L1 en reprenant le principe ayant servi à initialiser le réseau, c'est à dire que le poids des synapses pour le réseau L1 correspond à leur nombre d'activation multiplié par Ni, j. (notez que les synapses inactivées se retrouvent donc avec un poids nul, 0*Ni,j = 0).
On obtient alors le graphe de L1.
2) on procède de même pour L2, on obtient le graphe de L2.
3) le calcul de la distance entre L1 et L2 (DL1L2) est alors la valeur absolue de la différence des deux graphes.
Deux mesures peuvent alors être calculées,

(1) D1 L1- L2 = Σi,j | 1/Ni,j (L1i,j – L2i,j)|
et,
(2) D2 L1- L2 = Σi,j | Ni,j (L1i,j – L2i,j)|
La différence entre D1 et D2 s'explique ainsi, le poids attribué aux synapses dans la phase d'initialisation du réseau est d'autant plus grand que la synapse a été activée de nombreuses fois, ce qui implique dans le calcul de la distance qu'une langue diffère davantage d'une autre si elle possède un segment utilisant ces synapses, nous appelons ce type de segments des segments 'simples'. Ceci est appliqué dans la mesure D2 puisque nous utilisons directement Ni,j. En revanche la mesure D1 qui utilise 1/Ni,j va éloigner davantage une langue qui possède des segments 'complexes'. La confrontation des deux distances montre évidemment des résultats différents. Nous étudions actuellement les implications relatives à l'utilisation de l'une ou l'autre de ces mesures.

3. Conclusion et perspectives

Nous avons testé nos mesures de distance sur un échantillon de langues contenues dans UPSID 451, il comprenait 14 langues, 13 niger-kordofaniennes et une langue tchadique.

Après avoir obtenu des matrices de distance pour D1 et D2 (une pour le système, une pour les voyelles et une pour les consonnes), nous les avons soumises à une étude statistique (Multidimensional scaling) afin d'obtenir une représentation de type topologique.
Les résultats montrent, par exemple pour les voyelles, un groupement des langues possédant dans leur système des caractéristiques secondaires (nasalité, longueur) ou ayant un système avec de nombreux timbres différents, par opposition aux langues ayant moins de timbres et aucune caractéristique secondaire. L'interprétation est moins aisée pour les consonnes du fait de la plus grande variabilité des systèmes consonantiques par rapport aux systèmes vocaliques. Au niveau du système phonologique dans son ensemble, une représentation utilisant la mesure D2 permet des regroupements basés sur le croisement des critères de "complexité consonantique" et de "complexité vocalique".

Nous allons poursuivre notre recherche afin d'obtenir une échelle de complexité plus fine qui tienne compte de la notion de système tout en conservant les pondérations déjà obtenues au niveau des traits.

A plus long terme, notre objectif est d'utiliser cette mesure de complexité à des fins diachroniques, en comparant la complexité des systèmes synchroniques (UPSID 451) avec celle des systèmes reconstruits (BDPROTO 101, cf. Marsico, 1999), ce qui permettra une analyse qualitative de l'évolution des systèmes phonologiques : complexification, simplification ou stabilité.