La comparaison de typologie traditionnelle et de typologie phonolexique, basée sur la méthode des *n*-grammes, dans les dialectes baltes

Ansis Ataol Berzinch

Université de Lettonie,
Institut de Mathématiques et Informatique,
Boulevard de Rainis 29, Riga, Lettonie, LV-1459
Télephone: +371 7210428 - Télecopie: +371 7210128
Courrier électronique: *ataols@latnet.lv*

La question de la proximité des langues a toujours été une occasion de spéculations politiques et sociales. Bien sûr, il ne s'agit pas des langues sœurs ou des langues qui ont seulement quelques différences lexicales minimales; il s'agit des domaines formés par des langues apparentés, sans être trop proche. C'est pourquoi, depuis longtemps, nous avons eu l'idée trouver une procédure ou une série de procédures pour l'estimation automatisée de la proximité des langues.

Il faut reconnaître que cette idée n'est pas nouvelle. Par exemple, Yuri Tambovtsev compte les distances phonotypologiques entre langues à l'aide de la formule de la distance euclidienne dans l'espace à huit dimensions (ces dimensions sont les fréquences d'occurrence des espèces différentes de consonnes) [6]. Mais ici nous voulons présenter une autre méthode que nous avons appliquée avant de prendre connaissance des œuvres de Tambovtsev.

En 1994, William Canvar et John Trenkle ont proposé d'utiliser des listes de fréquence des *n*-grammes[1] pour la catégorisation du texte [5]. Il est vrai qu'en appliquant la loi de Zipf, il est possible de ranger la quasi-totalité de mots (dans notre cas, des groupes de signes ou des groupes de sons) par la fréquence d'usage. Canvar et Trenkle proposent de dresser des listes de fréquence des *n*-grammes pour divers textes et, par la comparaison de ces listes avec la liste de fréquence des *n*-grammes du texte introduit, de déterminer la catégorie à laquelle cela se rapporte. De cette façon, il est possible de déterminer automatiquement la langue, le codage et même le sujet du texte. Qu'y a-t-il de commun avec notre problème ? Le fait est qu'en comparant des listes de *n*-grammes, un nombre caractérisant le degré de ressemblance des textes est calculé[2]. Il est donc intéressant de savoir s'il est possible d'appliquer cette méthode pour déterminer de la ressemblance entre les langues.

Ceci est évidemment possible seulement dans le cas d'un usage de l'écriture phonétique uniforme, puisque des traditions différentes d'écriture et de codage, qui aident à trouver la décision pour le problème de catégorisation, rendent notre problème absurde. En outre, les exigences envers la catégorisation sont beaucoup plus faibles que les exigences envers la détermination de la proximité, c'est pourquoi nous pouvons supposer que pour obtenir des résultats satisfaisants, nous devrons travailler avec des textes beaucoup plus volumineux.

Les normes actuelles pour l'introduction des données phonétiques dans les ordinateurs ne nous arrangent pas: ou bien les normes connues (diverses normes locales, Unicode, SAMPA, IPA/ASCII) ne contiennent pas tous les sons, ou bien chaque son est décrit par un nombre de symboles (1 ou 2 octets). Il est beaucoup plus confortable de travailler sur des textes où chaque son serait codé sur un même nombre d'octets. Puisqu'il ne suffit pas d'un octet (256 permutations) pour la description de tous les phonèmes des langues de l'humanité, il est nécessaire de les coder sur au moins deux octets (65536 permutations). Certes, une division plus détaillée est théoriquement possible, mais nous paraît inutile, car elle dépendrait trop de la perception individuelle des transcripteurs. En outre, autant que nous sachions, il n'y a pas de normes en vigueur qui utiliseraient une transcription plus détaillée. Évidemment, il serait raisonnable dans le futur d'élaborer des polices spéciales Unicode, ainsi qu'un éditeur et des convertisseurs adaptés. Pour le moment, pour travaille sur des transcriptions phonétiques des langues baltes, nous nous sommes limités à un pseudo-code sur deux octets: chaque son rencontré est décrit par deux symboles ASCII selon un schéma défini. Malheureusement, d'après ce que nous savons, des transcriptions phonétiques des langues baltes au format électronique ne sont actuellement pas disponibles. C'est pourquoi nous devons introduire dans l'ordinateur des recueils imprimés existants [2], [3], [4]. Limités par le temps, nous nous sommes cantonnés à 4 modèles des patois de la Lettonie (3 latgaliens et 1 courlandien), de 500 jusqu'à 1000 symboles phonétiques. Ensuite, nous avons modifié le programme TextCat (écrit en PERL par le Hollandais Gertjan van Noord en 1994, suite à la publication de Canvar et Trenkle) pour pouvoir travailler avec des symboles de deux octets au lieu d'un octet, i.e., des *n*-grammes composés de $2n$ octets au lieu de n octets.

[1] Dans le cas présent, le terme «*n*-grammes» désigne les sous-chaînes (des mots du texte) de longueur n.

[2] La somme des différences des nombres dans l'ordre de listes si la *n*-gramme est présent, et des longueurs des listes sinon.

Ci-dessous, nous présentons les résultats de notre comparaison.

Tab. 1 – *Résultats de la comparaison.*

Langue du texte introduit	*Langue du texte comparé*	*Distance conventionnelle entre le texte introduit et le texte comparé*
Chkylbani (Škylbāni)	Baltinova	118344
	Nierza	125831
	Djukste	128467
Baltinova (Baļtinova)	Chkylbani	118336
	Nierza	118369
	Djukste	130354
Nierza (Nierza)	Baltinova	118353
	Chkylbani	125811
	Djukste	133562
Djukste (Džūkste)	Chkylbani	128501
	Baltinova	130396
	Nierza	133584

Dans la table 1, nous voyons que, malgré un volume exceptionnellement petit de textes en transcriptions phonétiques introduits, les résultats de notre expérience sont conformes à nos attentes. Ainsi, deux patois des communes voisines (le baltinovien et le shkylbanien) sont les plus proches l'un de l'autre. Viennent ensuite un patois latgalien (le nierzien), est et le djukstien de la Courlande .

Pour la nerzien, le plus proche se trouve être le baltinovien, ce qui est fondé aussi bien géographiquement que linguistiquement (*cf.* figure 1). Mais le plus éloigné, bien sûr, est le djukstien. Le plus proche des patois latgaliens pour le djukstien se trouve être le shkylbanien, ce qui n'est pas étonnant: les patois de Latgalie du Nord sont plus proches du djukstien aussi bien du point de vue lexical que du point de vue de la prononciation des formes grammaticales particulières.

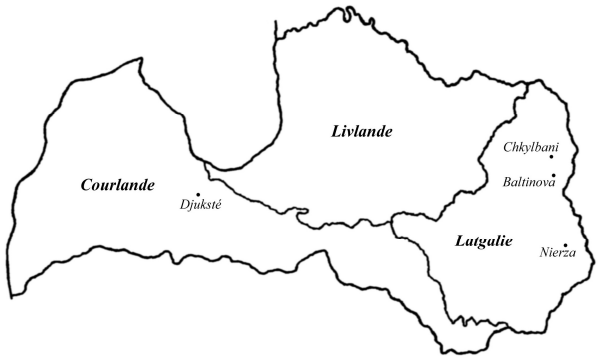

Fig. 1 – *Carte de Lettonie.*

Ainsi, à partir de textes peu volumineux et d'une procédure binaire de comparaison[1], nous avons obtenu des résultats relativement satisfaisants. Cela nous permet de supposer qu'en utilisant des transcriptions phonétiques plus volumineuses et de meilleure qualité, ainsi qu'une approche plus scrupuleuse dans l'élaboration de la procédure de comparaison, les résultats peuvent être très bons. Sur cette base, nous serons capables de faire d'authentiques conclusions. Pour continuer les recherches dans la direction entamée, il est nécessaire, entre autre:

- d'élaborer une table de codage sur deux octets pour les symboles phonétiques décrivant tous les phonèmes possibles des langues de l'humanité ;
- de créer pour ce codage des polices, un éditeur de texte et des convertisseurs pour d'autres normes de phonétiques ;
- de transcrire dans ce codage des corpus des patois de différentes langues du monde (en particulier pour nos expériences dans les domaines balte et slave) ;
- d'introduire dans l'espace des phonèmes une mesure définissant la distance phonétique pour chaque paire de ponèmes, fondée sur les propriétés acoustiques des sons et les particularités physiologiques de leur prononciation ;
- d'élaborer une procédure de comparaison des langues, basée sur cette mesure.

Références

[1] M. Rudzīte. *Latviešu dialektoloģija*. Rīga: Latvijas Valsts izdevniecība, 1964.

[2] M. Rudzīte. *Latviešu izlokšņu teksti*. Sast.. Rīga: P. Stučkas Latvijas Valsts universitāte, 1963.

[3] *Latviešu izloksnes. Augšzemnieku dialekta teksti*. Sast. N. Jokubauska. Rīga: Zinātne, 1983.

[4] B. Laumane *Latviešu izlokšņu teksti*. Sast.. Liepāja: Liepājas Pedagoģiskā akadēmija, 2000.

[5] W.B. Canvar & J.M. Trenkle. «*N*-Gram-Based Text Categorization». In *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, UNLV Publications/ Reprographics, pp. 161-175, 1994.

[6] Y. Tambovtsev. «Phonological Similarity Between Basque and Other World Languages Based on the Frequency of Occurrence of Certain Typological Consonantal Features». *The Prague Bulletin of Mathematical Linguistics, 79-80*, pp. 121-126, 2003.

[1] Nous comparons les lignes des listes des fréquences des *n*-grammes par le principe binaire (i.e. coïncidant: 0 ; ne coïncident pas: 1).