MATHÉMATIQUES ET SCIENCES HUMAINES

DAVID SANKOFF

Le vocabulaire partagé par des sous-groupes d'une communauté

Mathématiques et sciences humaines, tome 121 (1993), p. 41-47

<http://www.numdam.org/item?id=MSH_1993__121__41_0>

Math. Inf. Sci. hum. (31e année, n°121, *1993*, p. 41-47)

LE VOCABULAIRE PARTAGÉ PAR DES SOUS-GROUPES D'UNE COMMUNAUTÉ

David SANKOFF[1]

RÉSUMÉ — *On propose un indice de vocabulaire partagé γ afin d'évaluer les ressemblances et les différences entre les ensembles de mots utilisés dans deux sous-groupes d'une communauté. Cet indice mesure la différence entre le nombre moyen de mots partagés par deux locuteurs, l'un dans le premier groupe, l'autre dans le deuxième et le nombre prédit par une hypothèse nulle basée sur une distribution globale de la fréquence des mots. La formulation de γ permet des variations dans la taille de l'échantillon lexical d'un locuteur à l'autre. On présente la formule pour la variance de γ sous l'hypothèse nulle. Une application de l'indice à des données sur les emprunts à l'anglais dans le français parlé à Ottawa-Hull nous aide à comprendre l'utilisation de ces emprunts à l'intérieur des différents groupes d'âge.*

SUMMARY — The shared vocabulary of two subgroups in a community.
An index of sharedness γ is proposed for evaluating how similar or different are the lexical stocks of two subgroups of a larger comunity. This index measures the average number of words common to the vocabulary of two speakers, one in the first group, and on in the second, in excess of (or less than) the number predicted by a null hypothesis based on a global word-frequency distribution. The formula for γ alows the size of the vocabulary sample to vary from speaker to speaker. An expression is found for the variance of γ under the null hypothesis. Applying the index to data on borrowings from English in Ottawa-Hull French leads to an understanding of the use of loanwords among different age groups.

Comment évaluer quantitativement les différences lexicales entre deux ou plusieurs sous-groupes d'une communauté linguistique ? Cette question se pose souvent en sociolinguistique et des questions semblables sont soulevées en dialectologie et en linguistique historique. Nous explorerons ici une classe de modèles d'échantillonnage représentant le choix des mots utilisés par les membres d'une communauté et formulerons des tests statistiques sur les différences entre sous-groupes. Notre objectif principal est de tenir compte de l'éventualité d'une grande disparité de taille des échantillons d'une individu à l'autre. Ainsi on éviterait le biais dû à cette disparité lors de analyses statistiques.

Soit **L** l'ensemble de mots différents dans un corpus constitué d'un échantillon de langue parlée de M locuteurs, où $|\mathbf{L}| = n$. Dans certaines études, ce corpus est restreint à une classe de mots, telle que les substantifs, les mots grammaticaux ou les emprunts. Supposons que la taille du corpus (nombre total) de mots est N et que nous comptons $N_J(x)$ occurrences du mot x de l'ensemble **L** dans le discours du locuteur J, où J = 1,...,M.

1. Centre de recherches mathématiques, Université de Montréal.

Donc

$$N = \sum_{J=1}^{M} \sum_{x \in L} N_J(x).$$

Dénotons par

$$N_J = \sum_{x \in L} N_J(x)$$

le nombre total de mots dans le discours du locuteur J,

$$p(x) = \sum_{J=1}^{M} N_J(x)/N$$

la proportion du mot x dans le corpus,

$$p_J(x) = N_J(x)/N_J$$

la proportion du mot x dans le discours du locuteur J,

$$\mathbf{L}_J(x) = \{x \in \mathbf{L} \mid N_J(x) > 0\},$$

les mots utilisés par le locuteur J, où $n_J = |\mathbf{L}_J|$, et

$$\mathbf{L}_{IJ} = \{x \in \mathbf{L} \mid N_I(x) > 0 \text{ et } N_J(x) > 0\},$$

l'ensemble des mots utilisés en commun par les lecteurs I et J, où $n_{IJ} = |\mathbf{L}_{IJ}|$. Nous nous intéressons ici surtout au vocabulaire $\mathbf{L}_{IJ}$ partagé par deux locuteurs, en comparaison à $\mathbf{L}_I \setminus \mathbf{L}_{IJ}$ et $\mathbf{L}_J \setminus \mathbf{L}_{IJ}$, le vocabulaire utilisé exclusivement par l'un ou l'autre.

Nous voulons voir si ces ressemblances et différences lexicales sont attribuables aux regroupements socio-démographiques à l'intérieur de la communauté. Il y a des contextes où la variabilité des N_J ne joue guère. Dans ces contextes on peut étudier le degré de ressemblance des vocabulaires des locuteurs en adoptant des méthodes telle que l'analyse factorielle faite soit sur un tableau de 0 et 1 selon qu'un mot x apparaît ($N_J(x) > 0$) ou n'apparaît pas ($N_J(x) = 0$) chez un locuteur J, soit sur le tableau des $N_J(x)$ eux-mêmes. Par exemple, si tous les $N_J(x)$ non-nuls étaient assez grands, on saurait que le fait que $N_J(x) = 0$ n'est pas dû à la petitesse de l'échantillon (N_J), sinon à l'exclusion véritable du mot x du vocabulaire du locuteur J. Ou encore, si tous les N_J étaient égaux (= N/M), il n'y aurait aucune possibilité d'effet dû à la taille des échantillons.

En revanche s'il y a plusieurs mots pour lesquels $N_J(x) = 1$ ou d'autres valeurs basses, comme c'est généralement le cas, et qu'en plus les échantillons sont de taille N_J variable, situation tout autant répandue, des approches qui n'en tiennent pas compte pourraient être trompeuses. La taille n_{IJ} du vocabulaire commun est une variable aléatoire qui dépend de façon assez complexe de N_I, de N_J et des fréquences de mots dont la loi est inconnue. Ainsi le fait que n_{IJ} soit plus grand que N_{HK} pourrait être dû autant aux valeurs plus élevées de N_I et N_J par rapport à celles de N_H et N_K qu'à la plus forte ressemblance des vocabulaires des locuteurs I et J comparée à H et K. D'où l'intérêt pour l'étude de la relation entre N_I, N_J et n_{IJ}.

Considérons donc deux sous-ensemble de locuteurs A, B $\subseteq \{1,...,M\}$; où $A \cap B = \emptyset$. Si chaque $\mathbf{L}_J$ était tiré indépendamment de $\mathbf{L}$ selon une loi d'échantillonnage uniforme, avec les n_J fixés au préalable, on s'attendrait à ce que n_{IJ} soit approximativement égal à $n_I n_J / n$. Donc

$$\alpha(A,B) = (|A||B|)^{-1} \sum_{I \in A} \sum_{I \in A} (n_{IJ} - n_I n_J / n)$$

serait un indice sensible aux différences entre les locuteurs de A et de B, pris dans leur ensemble. Sous l'hypothèse nulle,

$$\mathbf{E}(\alpha) = 0$$

et, puisque l'échantillonnage est fait sans remise,

$$\mathrm{Var}(\alpha) = (|A||B|)^{-2} \sum_{I \in A} \sum_{J \in A} n_I n_J (n-n_I)(n-n_J) / n^2 (n-1),$$

si on simplifie en laissant tomber les covariances entre différentes paires de locuteurs. (Cette simplification a peu de conséquences si A et B sont grands. La variable N_{IJ} a une covariance *non-triviale* avec les variables N_{HK} de seulement $|B| + |A| - 2$ autres paires (H,K) de locuteurs, celles où H = I ou K = J, et a une covariance nulle avec $|B||A|-|B|-|A| + 1$ paires).

Si $\alpha(A,B)$ était significativement plus grand que zéro, cela appuierait l'hypothèse alternative selon laquelle il existe une forte association entre les deux sous-groupes. Le fait que l'indice soit significativement plus petit que zéro serait une évidence de l'éloignement entre A et B.

Remarquons que l'indice α est fonction du nombre de mots dans $\mathbf{L}_{IJ}$ et ne tient pas compte des p(x). Or, même si on ne s'intéresse qu'aux différences qualitatives entre A et B, c'est-à-dire aux mots qui sont utilisés et non pas aux fréquences de ces mots, un test de signification de l'indice α, où l'hypothèse nulle se base sur la loi uniforme, est peu approprié. On sait que les lois de fréquences de mots sont loin de la loi uniforme.

On pourrait, bien sûr, contourner ce problème en comparant les groupes de locuteurs selon un indice tel que

$$\beta(A,B) = (|A||B|)^{-1} \sum_{I \in A} \sum_{J \in B} \sum_{x \in \mathbf{L}} w[p(x)] \, |p_I(x) - p_J)|,$$

ce qui refléterait plus explicitement les différences entre les locuteurs au niveau statistique. Cependant, le choix des poids w[p], qui vise à compenser les énormes décalages entre les fréquences des mots répandus et celles des mots rares, reste toujours arbitraire (par exemple : w[p] = 1, w[p] = p ou w[p] = log p). De plus, du point de vue linguistique, les $\mathbf{L}_{IJ}$ et les n_{IJ} nous intéressent davantage que les statistiques d'ajustement de lois de répartition comme β.

Il serait donc souhaitable de retenir un indice du type α et de se servir d'une hypothèse nulle plus réaliste que la loi uniforme en ce qui concerne les échantillons $\mathbf{L}_J$. Pour ce faire, nous supposons que chaque $\mathbf{L}_J$ est construit non pas par tirage uniforme, mais selon les p(x). (D'ailleurs, nous considérerons dorénavant les p(x) comme étant des probabilités données, afin de simplifier l'analyse). Nous adoptons les hypothèses selon lesquelles les N_J sont fixés au préalable et que le tirage de $\mathbf{L}$ se fait *avec remise*.

La probabilité que le mot x soit dans $\mathbf{L}_J$ est donc

$$1 - q(x)^{N_J}$$

parce que $q(x) = 1 - p(x)$ est la probabilité qu'à chacun des N_J tirages le mot x ne soit pas choisi. Ainsi, l'indépendance nous assure que

$$E(n_{IJ}) = \sum_{x \in \mathbf{L}} (1-q(x)^{N_I})(1-q(x)^{N_J})$$

d'où nous obtenons l'indice

$$\gamma(A,B) = (|A||B|)^{-1} \sum_{I \in A} \sum_{J \in A} \left\{ n_{IJ} - \sum_{x \in \mathbf{L}} (1-q(x)^{N_I})(1-q(x)^{N_J}) \right\}$$

Soit $q(x,y) = 1 - p(x) - p(y)$. Sous l'hypothèse nulle,

$$E(\gamma) = 0$$

et, toujours en négligeant les covariances entre différentes paires de locuteurs,

$$\begin{aligned}
\text{Var}(\gamma) &= (|A||B|)^{-2} \sum_{I \in A} \sum_{J \in B} (E(n^2_{IJ}) - [(n_{IJ})]^2) \\
&= (|A||B|)^{-2} \sum_{I \in B} \sum_{J \in B} \left\{ \sum_{x \in \mathbf{L}} [1-q(x)^{N_I}][1 - q(x)^{N_J}] \right. \\
&\quad + \sum_{y \neq x \in \mathbf{L}} [1 - q(x)^{N_I} - q(y)^{N_I} + q(x,y)^{N_I}][1 - q(x)^{N_J} - q(y)^{N_J} + q(x,y)^{N_J}] \\
&\quad \left. - \left(\sum_{X \in \mathbf{L}} [1 - q(x)^{N_I}][1 - q(x)^{N_J}] \right)^2 \right\} \\
&= (|A||B|)^{-2} \sum_{I \in A} \sum_{J \in B} \left\{ \sum_{x \in \mathbf{L}} [q(x)^{N_I} + q(x)^{N_J} - q(x)^{N_I+N_J}][1 - q(x)^{N_I}][1 - q(x)^{N_J}] \right. \\
&\quad + \sum_{y \neq x \in \mathbf{L}} q(x,y)^{N_I+N_J} - q(x)^{N_I+N_J} q(y)^{N_I+N_J} \\
&\quad + [q(x,y)^{N_I} - q(x)^{N_I} q(y)^{N_I}][1 - q(x)^{N_J} - q(y)^{N_J}] \\
&\quad \left. + [q(x,y)^{N_J} - q(x)^{N_J} q(y)^{N_J}][1 - q(x)^{N_I} - q(y)^{N_I}] \right\}.
\end{aligned}$$

Connaissant la variance de γ sous l'hypothèse nulle, nous sommes en mesure de tester statistiquement la différence entre les groupes A et B.

Comme nous l'avons déjà mentionné, nous pouvons nous servir de la mesure γ pour comparer deux groupes de locuteurs à partir non seulement du vocabulaire entier **L**, mais aussi

de sous-ensembles de **L** tels que les noms, les verbes ou les mots qui commencent par une voyelle. Plus loin, nous allons discuter d'un exemple où γ sert à comparer le comportement des locuteurs francophones dans la région d'Ottawa-Hull par rapport à leur utilisation d'emprunts lexicaux à l'anglais. Nous signalons d'abord que lors de l'étude de ces emprunts, il s'est avéré que

$$\sum_{x \in \mathbf{L}} \left(1 - q(x)^{N_I}\right)\left(1 - q(x)^{N_J}\right)$$

prédit systématiquement des valeurs trop élevées pour n_{IJ}. Ceci est dû à l'hypothèse implicite du modèle quant à l'indépendance des N_J tirages qui constituent l'échantillon du locuteur J. En effet, il serait beaucoup plus raisonnable d'incorporer dans le modèle le fait que, une fois choisi par un locuteur, un mot a des chances considérablement augmentées de réapparaître une deuxième (troisième, quatrième...) fois dans son discours. La façon la plus simple serait de conserver les mêmes probabilités d'échantillonnage p(x) en adoptant l'hypothèse selon laquelle lorsqu'un mot x est choisi par un locuteur, r exemplaires de ce mot sont ajoutés à l'échantillon, où $r \geq 1$ est une variable aléatoire. Donc

$$N_J = \sum_{x \in L_J} r(x)$$

$$\approx \nu_J E(r)$$

où ν_J est le "vrai" nombre de choix indépendants faits en tirant l'échantillon du locuteur J. Dans ce cas,

$$\sum_{x \in \mathbf{L}} \left(1 - q(x)^{\nu_I}\right)\left(1 - q(x)^{\nu_J}\right)$$

$$\approx \sum_{x \in \mathbf{L}} \left(1 - q(x)^{N_I / E(r)}\right)\left(1 - q(x)^{N_J / E(r)}\right)$$

prédirait des valeurs plus appropriées pour n_{IJ}. Nous avons estimé $E(r) \approx 3/2$, comme la valeur qui minimise la somme des carrés des décalages entre les n_{IJ} et leurs valeurs prédites.

L'étude d'où proviennent nos données (Poplack, Sankoff et Miller 1988) est basée sur un corpus de 2,5 millions de mots du discours français enregistré chez 120 locuteurs adultes francophones répartis dans cinq quartiers des deux municipalité adjacentes de Ottawa, en Ontario, et Hull, au Québec. Les entretiens traitent de la vie quotidienne, dans le passé et le présent, des opinions sur les affaires courantes et de l'histoire personnelle des locuteurs. Nous avons repéré toutes les occurrences des mots d'origine anglaise dans la transcription informatisée du corpus et identifié comme emprunts tous ceux qui apparaissaient isolément dans le discours français (ni précédés ni suivis directement par d'autres mots d'origine anglaise) et qui fonctionnent normalement dans le phase comme un élément du français. Il y avait à peu près N = 20 000 occurrences et n = 2 000 emprunts différents. Nous avons distingué les emprunts établis et acceptés des emprunts "momentanés" ou "spontanés", mais cette distinction n'a pas de conséquences pour l'étude actuelle. Le cadre théorique de cette approche est esquissé par Poplack (1988).

Les locuteurs du corpus sont catégorisés selon l'âge, le sexe, le quartier de résidence, le niveau d'instruction, le type d'occupation et la compétence en anglais. Dans l'étude originale, nous avons calculé et présenté des valeurs de l'indice γ, et d'autres statistiques pour plusieurs de

ces facteurs ; le tableau ci-dessous montre, à titre d'exemple, les valeurs de l'indice γ (calculé selon l'ajustement $E(r) = 3/2$) pour comparer les différents groupes d'âge. Étant donné que, ignorant sa fonction de répartition, nous avons remplacé la variable aléatoire r par son espérance dans la prédiction de n_{IJ}, des tests de signification ne pourraient être qu'approximatifs. Néanmoins, puisque chaque groupe contient 20 locuteurs, donc $|A||B = 400$, et que les n_{IJ} prédits varient en général entre 10 et 20, nous sommes assurés que $|\gamma| \geq 0,7$ est significatif.

AGE	15-24	25-34	35-44	45-54	55-64	65+
15-24	- 1,6	- 1,2	- 0,7	- 0,5	- 1,4	- 2,0
25-34		0,0	0,2	0,6	- 0,2	- 1,0
35-44			0,7	1,3	0,8	- 0,3
45-54				2,1	2,3	0,6
55-64					2,0	0,7
65+						0,2

Nous nous dispensons de la condition $A \cap B = \emptyset$ pour le pas $A = B$. Donc, le calcul de γ implique, au lieu de $|A||B| = |A|^2$ termes, seulement les $|A|^2 - |A|$ termes où $I \neq J$. Ce calcul donne les chiffres sur la diagonale dans le tableau et mesure la diversité à l'intérieur du groupe A.

Nous voyons donc que le groupe des plus jeunes partage relativement peu d'emprunts avec les autres groupes. D'ailleurs, on avait déjà découvert que ce groupe manifeste une utilisation très diverse et innovatrice des mots anglais dans leur discours français. La même tendance s'observe chez le deuxième groupe, mais de façon moins accentuée. Les plus âgés ne partagent pas non plus un grand nombre de mots avec les autres groupes. Cet effet est dû au non-usage de mots archaïques chez les plus jeunes. Ce sont les autres groupes de locuteurs, notamment ceux âgés de 45-54 ans et de 55-64 ans, qui affichent les valeurs élevées pour les comparaisons intra-groupes et inter-groupes.

On peut déduire deux effets globaux du tableau. D'accord, chaque groupe tend à partager relativement plus avec ses voisins, ce qui s'explique par une communication plus fréquente entre les personnes d'à peu près le même âge ou bien simplement par le fait d'avoir acquis la langue (y compris les emprunts) à la même époque. La deuxième tendance est celle de l'innovation et de l'expérimentation lexicales chez les jeunes, par contraste avec la stabilité relative du vocabulaire de leurs aînés, et le caractère archaïsant d'une certaine proposition des emprunts utilisés par les plus âgés.

En conclusion, nous avons ici des préliminaires statistiques d'une analyse du vocabulaire partagé par différents sous-groupes d'une population. L'application de cette méthode à un ensemble important de données nous a permis d'arriver à des résultats intéressants. Pour améliorer l'analyse, il resterait à évaluer les simplifications implicites du modèle et du développement mathématique. Par exemple, nous avons considéré les p(x) comme des valeurs données en formulant l'hypothèse nulle. Il serait utile dans une analyse plus poussée d'incorporer l'erreur d'estimation implicite dans notre définition de p(x).

Nous avons choisi de ne pas tenir compte des covariances entre différentes paires de locuteurs, ce qui pourrait fausser le calcul des variances. Cependant, comme nous l'avons expliqué, cet effet sera minimal quand A et B sont grands.

Finalement, c'est dans l'hypothèse où l'échantillonnage de tous les $\mathbf{L}_J$ est régi par les mêmes p(x) que le modèle paraît le moins réaliste. Avant d'élaborer un modèle plus raffiné il faudra peut-être effectuer une analyse des $N_J(x)$ dans de grands corpus.

BIBLIOGRAPHIE.

POPLACK S., "Conséquences linguistiques du contact des langues : un modèle d'analyse variationniste", *Langage et société*, 43, 1988, 23-48.

POPLACK S., SANKOFF, D., MILLER, C., "The social correlates and linguistic processes of lexical borrowing and assimilation" *Linguistics*, 26, 1988, 47-104.