Statistique de fréquence et grammaire comparée

Dans ses *Remarks on Linguistic Affinity* (*Ural-altaische Jahrbücher*, tome XXVII, 1—2, pp 1—6), notre éminent confrère Björn Collinder a évoqué les problèmes soulevés par la comparaison lexicale entre langues apparentées et il a cru devoir faire état des théories exprimées par l'Américain Swadesh.

Il n'est pas dans notre intention de revenir ici sur ces théories dont nous avons dit ailleurs tout ce qu'elles ont de contestable. Il nous paraît cependant utile de considérer le problème à la lumière de données qui sont à notre disposition depuis que s'est achevée l'enquête effectuée en vue de déterminer les éléments les plus fréquents du lexique français.

On trouvera tous les renseignements concernant cette enquête dans l'ouvrage intitulé *L'élaboration du français élémentaire*, par G. Gougenheim, R. Michéa, P. Rivenc et A. Sauvageot (Paris, Didier, 1956).

La seule concession que nous ferons à Swadesh, c'est d'opérer avec les 200 vocables les plus fréquents du français. En faisant toutefois observer que cette fois-ci, c'est vraiment avec les 200 vocables les plus fréquents de la langue parlée que nous raisonnons. Il ne s'agit plus d'une liste établie arbitrairement mais d'une liste décelée par les enregistrements pris sur le vif. Pour la première fois en linguistique, nous disposons d'une liste de fréquence établie dans des conditions véritablement scientifiques et qui est donc susceptible de nous apporter des enseignements sûrs.

Or que relevons-nous? Que les 200 premiers mots de la langue française contemporaine (recueillis entre 1951 et 1954) comportent: 32 verbes, 21 substantifs, 6 adjectifs soit en tout 59 termes à valeur lexicale alors que les mots »grammaticaux» sont 141, soit plus de 70 % du vocabulaire de haute fréquence.

En regardant de plus près la composition des vocables »lexicaux», nous trouvons qu'ils ont tous des acceptions générales et parfois multiples. Voici au demeurant la liste des mots en question:

1) verbes : être, avoir, faire, dire, aller, voir, savoir, pouvoir, falloir, vouloir, venir, prendre, arriver, croire, mettre, passer, devoir, parler, trouver, donner, comprendre, connaître, partir, demander, tenir, aimer, penser, rester, manger, appeler, sortir, travailler.

L'ordre dans lequel ces verbes sont énumérés est celui de leur fréquence.

2) substantifs : heure, jour, chose, temps, fois, an, franc (monnaie), enfant, madame, maison, femme, gens, mois, soir, année, exemple, côté, matin, travail, histoire, voiture.

Cette énumération est également présentée selon l'ordre de fréquence.

3) adjectifs : petit, grand, bon, beau, vrai, vieux.

Les noms de nombre qui figurent parmi les 200 mots les plus fréquents sont : (dans l'ordre de fréquence)

deux, trois, cent, vingt, quatre, dix, cinq, huit, six.

On notera que neuf et sept ne viennent qu'après. Quant à onze et douze, ils traînent bien en arrière (aux 599-e et 607-e rangs, respectivement).

Le mot *un* étant article indéfini est très fréquent (14-e rang) mais comme numéral, il ne vient qu'au rang 217!

Le seul ordinal est premier (rang 165).

Ce qui est frappant, mais non inattendu, c'est que cette liste ne comprend pas de mots concrets proprement dits: ni termes désignant les parties du corps, ni vocables s'appliquant à la nourriture, au vêtement, à l'habitat (sauf *maison*). Même quand il s'agit de vocables désignant des objets, ils sont d'acception vague (*voiture*). En d'autres termes, il apparaît que les mots les plus fréquents ne sont que ceux dont l'acception est générale ou multiple. Les mots d'acception précise ne sont pas dénoncés par la fréquence. C'est qu'il s'agit d'une autre couche du lexique: celle des mots »disponibles» que les enquêteurs du *Français élémentaire* ont été contraints de faire »sortir» par d'autres procédés que ceux employés pour signaler les termes les plus fréquents.

Le lexique d'une langue se compose donc de trois sortes de mots: 1) des mots-outils ou mots grammaticaux, 2) des mots généraux de haute fréquence, 3) des mots précis de basse fréquence (mots disponibles).

Considérons maintenant le lexique sur lequel opère le comparatiste, plus particulièrement l'ouraliste. Les mots qu'il rapproche les uns des autres sont naturellement des mots grammaticaux (pronoms, mots auxiliaires, etc) et des mots le plus souvent de sens assez précis. Il en résulte que si l'on examine, à titre d'exemple, les mots comparés pour l'ouralien par Björn Collinder lui-même, dans son *Fenno-Ugric Vocabulary*, on trouve que l'énorme majorité est constituée par des mots »disponibles» dont les correspondants ne figurent pas dans la liste du Français élémentaire, certains ayant même un caractère technique très marqué.

On peut donc en induire que le lexique comparatif restitué par les étymologistes pour l'ouralien ou, plus restrictivement pour le finno-ougrien commun, est essentiellement fourni par des mots »disponibles» de fréquence rare ou quasi nulle.

Il y a à cela une raison. Les mots d'acception précise se laissent comparer plus évidemment entre eux. Le mot qui désigne le »père» ou la »mère», le »frère cadet» ou la »soeur cadette» ou tel animal ou telle plante ou tel instrument, tel ustensile, telle activité technique ou tel tour de main ne possèdent le plus souvent qu'une seule acception ou en tous cas un petit nombre d'acceptions. La teneur sémantique en est donc relativement simple. Or la comparaison est, en apparence du moins, d'autant plus sûre que l'acception sémantique qui se retrouve de part et d'autre est plus précise. C'est ainsi que le finnois *talvi* et le hongrois *tél*, désignant l'un et l'autre l'hiver, ne peuvent pas ne pas être identifiés, en dépit de la voyelle *a* du finnois qui jure avec l'*é/e* du hongrois. Dans l'enquête sur le français élémentaire, le mot »hiver» n'est venu en fréquence qu'au 535-e rang.

Cela revient à dire qu'en choisissant les 200 premiers mots d'une langue (dans l'ordre de fréquence), on exclut délibérément les termes qui se prêtent le mieux à la restitution des formes anciennes, ceux qui sont les plus sûrs parce que dépourvus d'ambiguïté sémantique.

C'est qu'il n'y a que deux points de départ possibles pour la comparaison: la ressemblance de forme phonique ou l'identité de contenu sémantique. Plus un mot est d'acception précise, plus son contenu sémantique paraît fournir un support solide pour la comparaison.

Cela ne veut pas dire que les mots de contenu sémantique précis et de forme ressemblante soient toujours à rapprocher, mais ils suggèrent les premiers la comparaison. C'est seulement par la suite que le linguiste rejette certains rapprochements obtenus par cette première confrontation. C'est ainsi que le hgr *csekély*, *sekély* »peu profond» a été longtemps associé au mot lapon *coakke* »lieu peu profond dans un cours d'eau» parce que les acceptions sémantiques étaient superposables et ne comportaient aucun aléa. C'est ici la forme des deux mots qui a fait rejeter finalement le rapprochement.

En partant de la forme phonique du mot, les erreurs sont d'autant plus fréquentes qu'on se satisfait de contenus sémantiques plus vagues ou plus divergents. C'est ainsi que le regretté Yrjö Wichmann avait à tort associé le zyriène *jur* »tête» (votiak *jjr*, etc »id») au finnois *järki* »raison, entendement», etc.

Le cheminement du comparatiste est donc sinueux: il compare entre eux des mots de sens voisin mais dont la forme peut présenter des dissemblances plus ou moins flagrantes, qu'il essaie d'expliquer, ou bien il confronte des mots d'aspect similaire mais dont les significations peuvent être passablement éloignées. Il essaie de rendre compte de ces divergences sémantiques en faisant état de changements sémantiques. Nous dirons volontiers que c'est en cela que consiste la »dialectique» comparative. Il s'agit de sortir d'un cercle vicieux en s'aidant de supports qui peuvent à chaque pas vous manquer. Mais les astronomes ne font pas mieux quand ils calculent le temps ou les distances entre les astres. A force de vigilance, d'ingéniosité et de précision, ils finissent par définir des résultats que l'expérience vérifie. On ne voit pas pourquoi les comparatistes se laisseraient décourager.

Toutefois, la nature même de cette investigation exclut que l'on opère uniquement avec les vocables les plus fréquents de la langue. Choisir les 200 premiers mots, je veux dire les 200

mots les plus usuels, c'est se condamner, en dehors des mots grammaticaux, à ne confronter d'une langue à l'autre que des termes d'acception sémantique trop vague et trop multiple pour fournir un appui solide à la détermination des correspondances phonétiques sans lesquelles il ne saurait y avoir de grammaire comparée.

Cela ne veut pas dire que les mots les plus fréquents d'une langue ne soient pas utiles à considérer. Au contraire, leur examen est indispensable.

D'abord parce qu'ils offrent les outils grammaticaux les plus usuels. Avec cette restriction, ou cette réserve si l'on préfère, que ces mots-outils sont plus ou moins nombreux selon la structure de la langue. En français contemporain, ils sont particulièrement nombreux du fait que la plupart des élargissements qui affectent les mots employés dans la phrase sont constitués par des mots-outils dont le sujet parlant a conscience qu'ils sont encore autonomes: pronoms personnels (*je*, *il*, *on*, *vous*, *elle*, *me*, *se*, *tu*, *nous*, etc, dans l'ordre de fréquence), démonstratifs (*ce*, *cette*, etc), des prépositions: (*de*, *à*, *en*, *dans*, *pour*, etc), des conjonctions (*que*, *mais*, *comme*, etc) sans parler des articles, etc. Or si nous nous reportons à l'ouralien, une partie de ces outils grammaticaux sont figurés par des suffixes, sans parler de ceux qui ne sauraient être restitués (conjonctions, relatifs, etc).

D'autre part, les mots figurant parmi les plus fréquents présentent volontiers des formes de type archaïque. C'est ainsi que les verbes les plus fréquents du français sont ceux qui gardent encore dans leur conjugaison des vestiges relativement bien conservés de l'état latin: *vous faites* (avec *-tes* à la 2-e pers. du pluriel au lieu de *-ez*).

Sans doute, si l'on pouvait comparer la liste des 200 mots les plus fréquents du français à celles de l'italien et de l'espagnol (quand elles seront établies à leur tour), on trouverait aisément que les correspondances entre les listes sont très nombreuses. Et les mots qui ne se retrouveraient pas de part et d'autre seraient néanmoins identifiés le plus souvent en faisant appel à des vocables plus rares ou d'acception différente. En pointant dans la liste française les mots d'origine latine (classique ou

tardive), on repérera à peu près l'ensemble des vocables. On constatera d'ailleurs la présence d'emprunts savants au latin (*exemple*, *comprendre*, *moment*, etc). A ce compte, les calculs établis par Swadesh sont totalement démentis. Les 200 premiers mots du français sont d'origine latine à une seule exception près (le mot *franc!*). Or le français de 1954 est quand même séparé du latin par au moins 15 siècles.

Ce résultat surprenant s'explique par les conditions où le français a évolué. Il est resté dans l'orbite latin et s'est rechargé de substance latine sous forme d'emprunts savants (*exemple*, *moment*, etc). Donc, les conditions dans lesquelles une langue vit se reflètent dans la constitution de son lexique de haute fréquence. Il est probable que les 200 mots les plus fréquents du roumain n'ont pas la même teneur de latinité. Ceci infirme également la théorie de Swadesh car la transmission ou la déperdition ne peuvent plus être évaluées dans les mêmes pourcentages.

A poursuivre notre examen, nous sommes amené à cette autre constatation que les mots les plus fréquents du français enferment entre autres les vocables *comprendre*, *penser*, *heure*, *franc* (monnaie), *histoire*, etc dont on a tout lieu de croire que les ancêtres des Finnois ou des Hongrois ne possédaient pas les équivalents. Sans parler du terme *moment*, tout moderne par ses implications. C'est que la fréquence reflète l'usage et les convenances de la société où se parle la langue. Les 200 mots les plus fréquents du hongrois de nos jours ou du finnois contemporain feraient apparaître d'autres termes typiques des moeurs des sociétés où se parlent ces idiomes.

A ces réserves près, il est visible que la comparaison de la liste française avec les listes parallèles non encore établies pour les autres langues romanes feraient apparaître quand même beaucoup de termes communs. Ces termes communs, à quelques emprunts près, qu'un oeil averti décèlerait aussitôt, attesteraient la parenté génétique des langues néo-latines.

Mais irait-on beaucoup plus loin? La comparaison de la liste allemande (non établie encore) et de la liste française ferait-elle ressortir une parenté? Partiellement seulement. Je veux dire que quelques mots outils se compareraient assez bien: *tu* ~ *du*,

me ~ *mich*, *non* ~ *nein*, *pour* ~ *für*, etc. Mais cette comparaison s'étendrait à combien de termes? Et elle ne ferait ressurgir aucun système, ni celui des pronoms (*nous*, *wir*, *uns*, *vous*, *ihr*, *euch*) ni celui des démonstratifs, des prépositions (*par*, *durch*, *dans*, *in*, etc). Quant au vocabulaire de valeur proprement lexicale, la discordance serait totale. En admettant même qu'on puisse, faute d'une liste allemande, comparer terme à terme les vocables de la liste française avec les mots allemands de signification correspondante, on aurait la confrontation suivante:

Pour les verbes:

être / sein
avoir / haben
faire / machen, tun
dire / sagen
aller / gehen
voir / sehen
savoir / wissen
pouvoir / können
falloir / müssen
vouloir / wollen
venir / kommen
prendre / nehmen, etc.

Certains équivalents seraient difficiles à définir: partir (? *abgehen*, *fortgehen*), sortir (*ausgehen*, *austreten*, etc). Certains rapprochements seraient suggérés à faux par une certaine similitude de forme: *avoir / haben*. Le seul rapprochement qui se dessinerait serait celui de *vouloir / wollen*. La parenté indo-européenne ne saurait être fondée sur des témoignages aussi disparates et aussi isolés. Même les noms de nombre prêteraient à difficulté: *trois* ~ *drei* et *six* ~ *sechs* se laisseraient rapprocher mais *vier* et *quatre*, *deux* et *zwei* (sans l'intermédiaire de l'anglais *two* par exemple); *fünf* et *cinq*?

En faisant état des formes de la conjugaison, on parviendrait à découvrir des »coïncidences» comme celle à laquelle notre maître Antoine Meillet attachait tant d'importance: *il*

est / er ist — ils sont / sie sind mais cela n'entraînerait guère au-delà. A moins de tenir compte de l'orthographe française qui rappelle l'existence d'un *-t* de 3-e personne du singulier (*il fait, il dit*, etc) ou d'un *-nt* de 3-e personne du pluriel (*ils font, ils disent, ils travaillent*). Pour le cas où l'on comparerait la forme purement orale du français contemporain à l'allemand également oral, de telles »évocations» feraient défaut. Surtout si l'on notait le français au moyen d'une transcription du genre de celle des *Finnisch-ugrische Forschungen!*

Encore s'agit-il d'opposer seulement le français et l'allemand qui, génétiquement, sont relativement proches l'un de l'autre aux yeux de l'indo-européaniste. Qu'adviendrait-il si nous ne possédions plus des langues indo-européennes que des vestiges plus éloignés les uns des autres, tels que le français et le russe.

Pour se rendre compte de la situation dans laquelle se trouverait alors le comparatiste, il suffira de confronter les 200 premiers mots du français tels que les a fait ressortir l'enquête sur le Français Elémentaire et les 204 mots les plus fréquents du russe tels qu'ils se présentent à la suite de l'investigation instituée par Harry H. Josselson (*The Russian Word Count*, Wayne University Press, Detroit 1953).

Sans doute, ces derniers termes ont été prélevés dans des textes écrits, leur fréquence a été établie d'après des règles discutables parce que purement arbitraires, mais il s'agit d'un sondage qui peut donner une idée des conditions dans lesquelles opérerait le linguiste disposant d'une véritable statistique de fréquence rigoureusement établie. Or cette statistique nous apprend que sur les 204 mots de »haute fréquence», on relève 40 substantifs, 12 adjectifs, 36 verbes et 10 noms de nombre (dont deux ordinaux), soit un peu moins de 50 % de mots »lexicaux». La proportion est donc nettement supérieure à celle attestée en français, ce qui provient en partie sans doute du caractère »flexionnel» du russe.

Parmi les mots dont l'acception correspond en gros à celle des mots relevés en français, on trouve:

1) verbes:

être	*byt'*,	*jest'*
faire	*delat'*,	*sdelat'*
dire	*skazat'*	
aller	*itti*,	*χodit'*
voir	*videt'*	*uvidet'*
savoir	*znat'*	
pouvoir	*moč́*	
vouloir	*χoteť*, etc.	

Le verbe *falloir* n'aurait pas de correspondant équivalent puisque le russe se sert de la construction avec *nado* pour exprimer la même pensée.

Les seuls verbes français auxquels pourrait s'accrocher une comparaison bien timide seraient: *voir* (*videt'*), *donner* (*dat'*, *davat'*). C'est peu.

2) substantifs:

heure	*čas*
jour	*den'*
chose	*delo*
fois	*raz*
an	*god*
maison	*dom*
gens	*ljudi*, etc.

Pas un des mots français ne ressemble aux substantifs russes de sens correspondant. Aucune comparaison ne pourrait être tentée.

3) adjectifs:

petit	*malenkij*
grand	*bol'šoj*
bon	*χorošij*, etc.

Ici non plus, aucune comparaison ne saurait être entrevue.

Quant aux noms de nombre, c'est tout au plus si l'on peut confronter directement *deux* / *dva*, *trois* / *tri* et à la rigueur *desjat'* / *dix*.

Le plus décevant est la comparaison des pronoms: *vous / vij*, *vam*, *vas*, *tu / ty*, *tebja* et à la rigueur *je / ja* et *me*, *moi / menja*, *mne*. Pourrait-on rapprocher *se* et *sebja*? Un linguiste un peu imaginatif le ferait sans doute, au risque de se faire critiquer. Quoiqu'il en soit, le système pronominal en tant que tel n'apparaîtrait pas à travers ces confrontations (*kto* ne pourrait être rapproché de *qui* ni *čto* de *quoi*, etc).

Que conclure de l'examen de ces faits?

Que la comparaison est une entreprise désespérée si elle doit se fonder sur une liste de mots limitée aux termes les plus usuels. La haute fréquence ne fait apparaître ni les correspondances de vocabulaire ni le système de la grammaire. Il est donc presqu'impossible a priori de retenir les éléments d'une comparaison tant que nous ne disposons pour tous matériaux que de listes de mots relevées par un enquêteur auprès d'un ou de plusieurs informateurs, à moins que cette liste ne comprenne quelques termes techniques ou un certain nombre de mots »disponibles» de sens aussi précis que possible.

Heureusement pour nous, les mots les plus faciles à recueillir ne sont pas ceux qui sont les plus fréquents mais ceux dont le sens est le plus concret. Les listes des enquêteurs ne comprennent même le plus souvent que ces termes de basse fréquence. C'est ce qui permet d'»accrocher» la comparaison.

Les éléments qui se prêtent le mieux à la comparaison sont les mots concrets, d'acception aussi étroite que possible et même les mots rares, désuets, qui ont conservé une forme archaïque. Ainsi le mot *húgy* »stella» du hongrois fournit un terme de comparaison avec d'autres mots ouraliens, voire ouralo-altaïques, alors que le terme usuel *csillag* ne trouve d'étymologie que par le biais d'études prolongées sur des éléments de vocabulaire disparates.

Mais les comparatistes n'avaient pas besoin de ces enseignements. Dès le début, ils ont considéré avec faveur les éléments rares et désuets du lexique comme aussi les formes rares et archaïques du système grammatical. C'est sur ces matériaux qu'ils ont construit l'édifice de la grammaire comparée. Cette tâche a été relativement facile quand on avait affaire à des langues attestées à date ancienne mais elle s'est avérée au

contraire très malaisée dès qu'il a fallu comparer des idiomes connus à date récente seulement. Les confrontations auxquelles nous venons de procéder nous enseignent en effet que les éléments les plus usuels d'une langue offrent peu de prise à la comparaison tant qu'on ne les confronte qu'aux éléments usuels d'un autre idiome. Les points de comparaison n'apparaissent que lorsque l'investigation s'étend aux parties les moins usuelles du lexique. C'est ainsi que le français *maison* ne saurait être raproché du russe *dom* mais par contre ce dernier mot dit »quelque chose» à tout Français qui songe au terme »domestique» (qui est d'ailleurs un emprunt au latin).

Or des difficultés de ce genre, nous nous y heurtons à chaque pas en essayant de rapprocher des mots ouraliens ou encore davantage si nous prétendons aller au delà de l'ouralien vers le youkaguir ou l'altaïque. Nos documents sont trop pauvres et ils sont trop récents.

Mais si nous songeons au peu de points communs relevés entre le français, l'allemand et le russe en opérant sur des listes restreintes bien que constituées par des mots fréquents nous devons penser que les déceptions rencontrées dans les efforts pour comparer l'ouralien, l'altaïque, etc ne signifient pas que ces idiomes soient sans affinité entre eux. Leur parenté est difficile à déterminer, faute d'éléments de démonstration suffisamment nombreux et suffisamment clairs mais rien n'autorise à induire que les liens unissant entre eux toutes ces langues soient moins authentiques et moins intimes que ceux qui unissent le français, l'allemand et le russe.

Nous irons même plus loin: à tout prendre, les éléments de comparaison entre les langues ouraliennes, altaïques, etc forment un ensemble nettement plus marqué. Les systèmes pronominaux, par exemple, se détachent dans leurs linéaments essentiels. Notre confrère Björn Collinder avait donc raison d'insister sur leur aspect »significatif». Le hasard pur ne ferait pas aussi bien les choses. En tout cas, il a omis de les faire aussi bien en ce qui concerne les 200 mots les plus usuels du français opposés à leurs équivalents allemands et russes.

Compte tenu des difficultés rencontrées dans la comparaison des langues ouralo-altaïques, il faut donc reconnaître que

cette comparaison se présente sous un aspect relativement favorable. Puisque nous opérons sur des éléments peu nombreux et récents, recueillis dans des conditions souvent précaires, l'ensemble des faits observés n'en acquiert que plus de force probante. Il n'est pas du tout absurde de vouloir rapprocher les langues ouraliennes du youkaguir, contrairement à ce que le regretté Paasonen avait trop hâtivement conclu. Il n'est pas non plus contraire à tout esprit scientifique de vouloir trouver quelque parenté aux langues ouraliennes et à celles dites altaïques. De glorieux précurseurs, comme M. A. Castrén n'avaient pas été trompés par leur instinct. Ils avaient tout de suite perçu les affinités qui suggéraient cette comparaison. Ils ont eu raison de formuler cette hypothèse.

Ou bien alors, il n'est qu'une autre théorie pour rendre compte de ces »similitudes»: celle d'une sorte de parallélisme glottogonique, une sorte d'ologénèse des langues. Les mêmes systèmes linguistiques ou des systèmes analogues auraient surgi indépendamment les uns des autres. Les mêmes phonèmes ou des phonèmes voisins auraient été chargés des mêmes fonctions ou de fonctions similaires. Mais alors on ne saisirait pas pourquoi des similitudes de ce genre ne se retrouvent pas entre un nombre plus grand de langues. On ne rendrait plus compte de la diversité des structures linguistiques. Pour avoir refusé d'accorder une valeur significative à des similitudes bien définies, on se résignerait à admettre une explication de l'évolution des langues qui nous ramènerait aux plaisanteries d'un Marr ou aux bévues d'un Trombetti. Les enseignements de la statistique de fréquence nous autorisent à penser que tout n'est pas perdu pour la grammaire comparée, même quand elle opère dans des conditions presque désespérées.

AURÉLIEN SAUVAGEOT