


Les distances linguistiques et leurs effets sur les comportements économiques*

VICTOR GINSBURGH

 VICTOR GINSBURGH est professeur honoraire de science économique. Il a enseigné en Belgique, en France et aux Etats-Unis. Il s'est intéressé tour à tour à l'équilibre général, à l'organisation industrielle, à l'économie de l'art et aux liaisons entre langues, culture et économie. Il est notamment co-auteur (avec Shlomo Weber) de *How Many Language Do We Need. The Economics of Linguistic Diversity*, Princeton University Press, 2011 et co-éditeur (avec Shlomo Weber) du «Palgrave Handbook of Economics and Language» à paraître en février 2016.

Email: vginsbur@ulb.ac.be //

Site internet : http://ecares.org/index.php?option=com_comprofiler&task=userProfile&user=112&Itemid=263

Résumé

Après avoir défini différentes manières de calculer les distances entre les langues (méthode lexico-statistique, distances de Levenshtein, distances basées sur les arbres linguistiques ou encore distances phonétiques) et entre groupes de population, ce chapitre propose d'explorer comment ces distances linguistiques ont été récemment utilisées par les économistes et permettent d'améliorer notre compréhension de certains comportements macro et microéconomiques dans des domaines tels que celui de la mesure de la diversité des populations, du commerce international, des migrations, de la traduction littéraire, de l'apprentissage des langues, ou encore des résultats du concours Eurovision. Ce chapitre montre également comment les distances linguistiques peuvent aider à analyser de réformes linguistiques, notamment dans la gestion des langues dans un pays ou une région où coexistent plusieurs langues.

Code JEL : F14, F22, J15, P00, Z11, Z13, Z19

Mots clés : diversité humaine, distances entre langues, commerce international, migrations, apprentissage des langues, traductions

*Des parties de ce chapitre sont basées sur Ginsburgh and Weber (2011 ; 2015). Merci à C. Carrère, J. de Melo et S. Weyers pour leur relecture attentive du manuscrit. A paraître dans *L'impact économique des langues - implications pour la francophonie* (2016), Carrère C. (dir.), Ferdi, Economica, Paris. Ce travail a bénéficié du financement de la Ferdi et d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence "ANR-10-LABX-14-01".

1. Introduction

Les recherches sur les distances entre les langues ne sont pas nouvelles. Les arbres linguistiques sont (presque) aussi vieux¹ que les arbres généalogiques construits par nos arrières grands-parents (mais toujours à la mode) pour remonter dans le temps et retrouver leurs (et nos) ancêtres. Les arbres linguistiques procèdent à peu près de la même façon. En partant des mots (de leurs sonorités ou de leur prononciation) tels qu'ils existent aujourd'hui on essaie de remonter étape par étape aux origines de nos langues. L'essai le plus spectaculaire est celui de Merritt Ruhlen (1997) qui pense avoir (re)trouvé 27 mots de la première langue parlée par l'humain il y a 50.000 ou 100.000 ans². Et nous verrons plus loin, que ces arbres peuvent servir de base à construire des distances, mais il existe également d'autres manières plus subtiles basées sur certaines des caractéristiques individuelles des langues telles que le vocabulaire, la syntaxe, la phonétique, etc. et qui devraient permettre de juger de leur parenté.

Les distances entre langues ont fait et font toujours l'objet de recherches par un grand nombre de linguistes qui s'intéressent aux apparentements entre les quelques 6.000 langues qui existent aujourd'hui. Ces recherches permettent de remonter dans le temps vers les langues qui ont engendré celles qui existent, et de dater les moments de leurs séparations. De fait, depuis longtemps, linguistes, généticiens et archéologues font œuvre commune dans les recherches sur nos ancêtres (voir par exemple Cavalli-Sforza, 2000 et Renfrew, 1990). L'intérêt des économistes s'est manifesté beaucoup plus récemment, et plus particulièrement depuis l'établissement de l'Atlas Narodov Mira (1964) qui a établi un classement ethno-linguistique des peuples de la terre qui ont été utilisés pour calculer des indices de diversité (ou de fractionalisation), dont certains tiennent compte des distances entre langues³. On s'est aussi aperçu aussi que les distances entre langues viennent en aide pour comprendre bon nombre de comportements économiques.

Ce chapitre comporte quatre parties. Les parties 2 et 3 sont consacrées aux éléments de la construction des distances entre *couples* de langues et à l'analyse de la distance linguistique *moyenne* entre pays et régions où des groupes de citoyens parlent des langues différentes. La quatrième partie s'intéresse à quelques applications

¹ Hombert et Lenclud (2014, p. 84) attribuent la naissance des recherches modernes au 2 février 1786, où dans une conférence, William Jones (1746-1796) « formule, bien plus clairement qu'auparavant, l'existence d'une relation entre le sanskrit, le grec, le latin, le celtique, le gothique, l'ancien persan [et] donne une explication de cette relation par la parenté entre ces langues : elles dérivent d'une même source et cette communauté d'origine est attestée tant par l'examen des 'racines verbales' que par celui des 'formes grammaticales' ». Le travail a été poursuivi par Franz Bopp (1791-1867), un des frères Grimm, Jacob (1785-1863), Karl Wilhelm Friedrich von Schlegel (1772-1829) entre autres.

² Pour autant qu'il y ait eu monogénèse, ce qui est contesté par certains chercheurs, dont Hagège (1996), qui ne remet pas en cause la monogénèse de l'espèce humaine, mais défend l'idée que plusieurs langues, non nécessairement reliées les unes aux autres, ont pu apparaître plus ou moins en même temps. Voir aussi Hombert et Lenclud (2014, p. 511 sq.).

³ Voir par exemple Desmet et al. (2009, 2012, 2015).

économiques dans lesquelles les distances linguistiques sont utilisées comme variables permettant d'améliorer l'explication certains comportements micro- ou macroéconomiques, tels que le commerce international entre pays, les migrations, la traduction littéraire, l'apprentissage des langues, et les résultats du concours Eurovision. La dernière partie montre comment les distances linguistiques peuvent être utilisées pour analyser des réformes linguistiques en simulant les effets d'une réduction ou d'une augmentation du nombre de langues officielles.

2. Distances entre couples de langues

Les langues se différencient (et se ressemblent) de diverses façons. Le vocabulaire est souvent tenu pour essentiel, et il est vrai que si l'anglais et l'allemand font partie d'une famille dite indo-européenne et de la branche des langues germaniques, les mots *moon* (lune en anglais) et *Mond* (lune en allemand) sont clairement reliés et se révèlent bien différents du mot *lune* en français, bien que le français, une langue romane, fasse également partie des langues indo-européennes. Mais même les mots *moon* et *Mond* ont des genres différents : le premier est neutre, le deuxième masculin et *lune* est féminin. La prononciation de *moon* est proche de *Mond*. Le *e* final de *lune* est muet, tandis que le *a* final dans le mot espagnol ou italien *luna*, ne l'est pas. Les diphtongues, c'est-à-dire les sons composés de deux ou plusieurs voyelles qui forment un seul son ajoutent également des distinctions. Ainsi l'étudiant français qui apprend l'anglais se demandera bien pourquoi la diphtongue *ou* de *south* (sud) ne se prononce pas comme celle de *tour* (qui vient sans doute du mot français *tour*, lorsqu'il décrit par exemple le tour de la ville). La phonétique contribue donc aussi à la distance entre deux langues et la manière dont les lèvres, la langue (celle qui se trouve dans notre bouche), les dents et les cordes vocales contribuent à produire les sons qui sont distinctifs des langues que nous parlons. Chaque langue contient des phonèmes que des locuteurs d'autres langues trouvent imprononçables. Il est, par exemple, plus facile pour un locuteur d'une langue qui contient un grand nombre de phonèmes d'apprendre une langue qui en contient peu, que l'inverse.

La syntaxe, qui étudie les relations entre les mots et leur ordre pour former des phrases, est une autre difficulté qui illustre les différences entre les langues. Comparez les trois phrases *j'aimerais observer la lune*, *ich möchte den Mond beobachten* et *I would like to observe the moon*. L'ordre des mots est différent entre l'allemand et l'anglais, deux langues germaniques, mais il est identique entre le français (langue romane) et l'anglais. Il est souvent difficile de comprendre rapidement le nombre *fünfundzwanzig* (vingt-cinq en allemand, mais qui se dit *cinq* et *vingt*) pour un francophone habitué à vingt-cinq, qui se dit cependant de la même façon, *twenty five* en anglais, une langue germanique comme l'est l'allemand.

Les grammaires constituent un casse-tête supplémentaire. L'allemand et le russe ont des déclinaisons ; il en reste une trace en anglais pour le génitif (on écrira par exemple *the moon's last quarter*, ce qui est un rien plus élégant que *the last quarter of the moon*) et le français a perdu les déclinaisons qui lui venaient pourtant en ligne presque directe du latin. L'étudiant français qui n'a étudié ni grec ni latin au lycée ne connaîtra probablement pas le mot *déclinaison* lui-même et encore un peu moins sa signification.

Ces quelques exemples simples illustrent combien est complexe le calcul d'un nombre unique qui représente la distance entre deux langues, sans même entrer dans la question plus fondamentale : les langues ont-elles une structure commune ?

La linguistique historique

La *linguistique historique* (ou *comparée*) qui consiste à trouver « un ancêtre commun, et à suivre sa descendance à travers le temps, tout en admettant une divergence graduelle de la source commune » (McMahon et McMahon, 2005, p. 3) n'est pas intéressée par la communicabilité entre locuteurs qui utilisent aujourd'hui des langues différentes. Elle vise à démontrer l'existence de relations entre des langues et la reconstruction hypothétique d'un ancêtre commun en utilisant des ressemblances morphologiques, les lexiques ou vocabulaires des langues, la syntaxe et les correspondances de sons, et identifie des groupes et des sous-groupes de langues suivant leur similitude (McMahon et McMahon, 2005, p. 5). Le résultat final est souvent présenté sous forme d'un arbre (linguistique) qui contient une racine (le langage ancestral), des branches qui elles-mêmes se subdivisent en branches plus fines, pour finir par des « brindilles » que sont les langues que nous utilisons aujourd'hui⁴. Ce résultat est comparable à ce que font les biologistes pour construire les arbres qui décrivent l'évolution des animaux. Pour engendrer un tel arbre, on commence par dresser un tableau qui contient les espèces animales (auxquelles on pense pouvoir trouver un ancêtre commun) ainsi que les caractéristiques qui les décrivent. Par tâtonnement (aujourd'hui, c'est l'ordinateur qui tâtonne en utilisant un algorithme), on trouve l'arbre considéré comme étant le « meilleur ». Cette méthode (aussi appelée de « comparaison de masse ») peut être, on s'en doute, plus ou moins rigoureuse⁵, et faire l'objet de controverses. Comme le soulignent McMahon and McMahon (2005), il est impossible de la soumettre à des tests statistiques rigoureux et il n'est pas exclu que deux groupes de chercheurs qui travaillent indépendamment aboutissent à des résultats différents.

⁴ Voir le Tableau 1 qui présente un arbre très simplifié des langues indo-européennes. Pour un exemple de l'arbre indo-européen « complet », voir <http://www.ethnologue.com/subgroups/indo-european>.

⁵ Voir McMahon et McMahon (2005, pp. 19-29) pour les détails.

Il convient également de noter que des recherches ont été menées sur les époques où les séparations des branches se sont faites. C'est l'objet de la *glottochronologie* que Swadesh (1972) avait basée sur l'hypothèse (controversée) que la probabilité que les mots perdent ou changent de sens est constante dans le temps. Le concept a cependant repris vie avec les articles de Gray et Atkinson (2003) et de Searls (2003). Ainsi, selon Gray et Atkinson (2003), le groupe indo-européen se serait séparé en « indo » (dont dérive l'hindi actuellement parlé en Inde) et en « européen » (dont dérivent les langues européennes) il y a quelque 7 000 ans, alors que la séparation entre le suédois et l'allemand daterait d'il y a moins de 2 000 ans.

Les distances basées sur les arbres linguistiques

Fearon et Laitin (1999), Laitin (2000) et Fearon (2003) ont proposé d'utiliser les distances entre les branches d'arbres linguistiques mis à notre disposition, par un groupe de linguistes qui œuvrent à un excellent outil de travail, *Ethnologue* (2009)⁶. Cette approche a deux avantages par rapport à celles qui suivront : (a) elle prend en compte les divers aspects qui caractérisent les (familles ou groupes de) langues tels que lexicque, syntaxe, grammaire, etc. et (b) les arbres (la méthode la plus anciennement étudiée) ont le mérite d'exister pour les quelque 6.900 langues⁷ connues aujourd'hui et répertoriées par *Ethnologue*. Ces distances s'avèrent cependant moins précises que les distances lexicostatistiques dont il sera question plus loin.

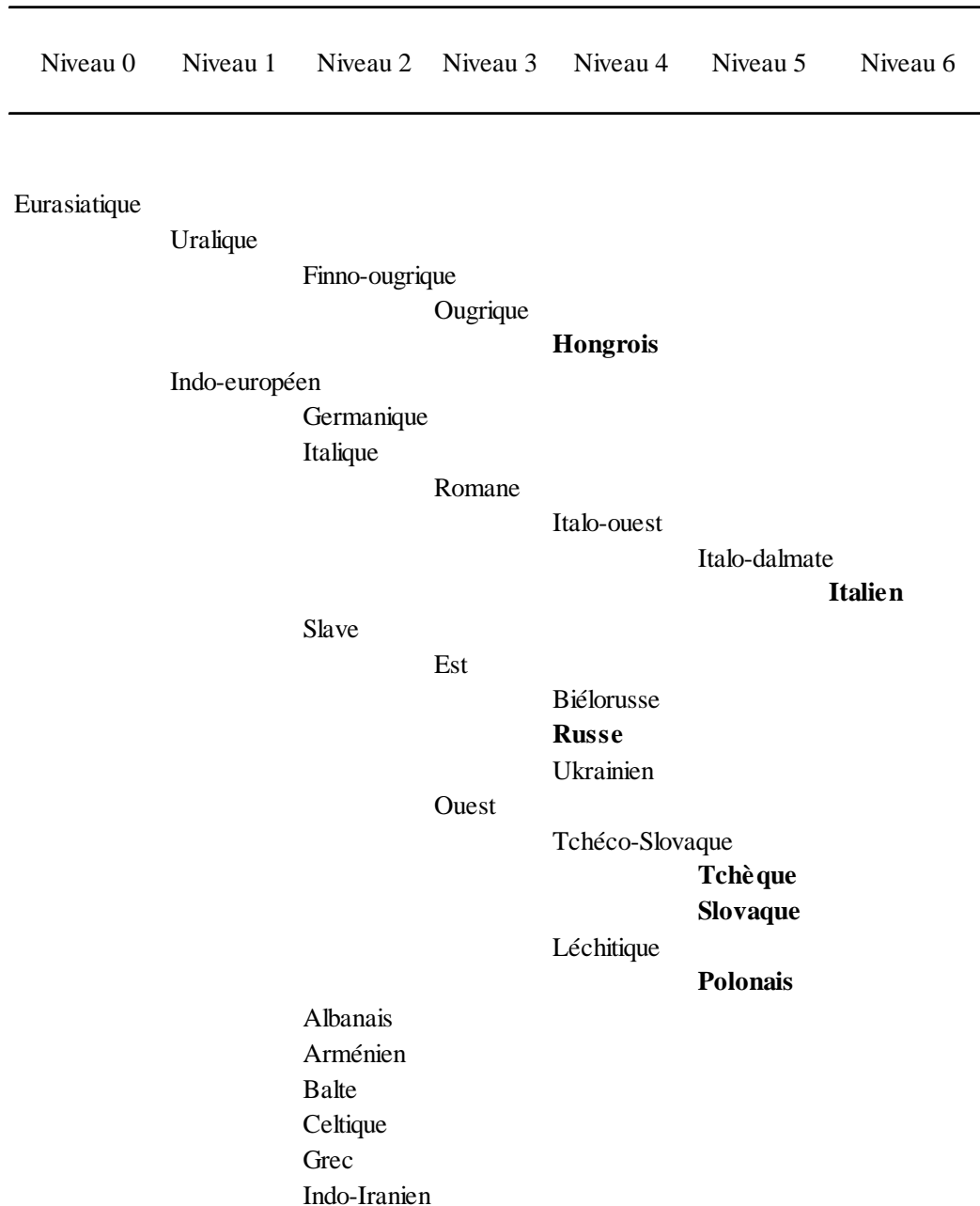
L'exemple que nous utilisons pour décrire la méthode est basé sur les langues en caractères gras qui figurent dans le Tableau 1. Nous comparons ici la distance entre le tchèque et les autres langues en caractères gras (hongrois, italien, russe, polonais et slovaque). Le tchèque et le hongrois proviennent de deux familles linguistiques qui n'ont pas de relation entre elles (si ce n'est qu'au niveau 1 de l'arbre, elles sont toutes deux eurasiatiques). Le tchèque est une langue indo-européenne, le hongrois fait partie des langues ouraliques, qui se séparent dès la première branche. Leur rang est 1. Le tchèque et l'italien partagent le fait d'être deux langues indo-européennes, mais se séparent au niveau 2 en italique et slave, ce qui leur donne un rang 2. Le tchèque et le russe sont des langues slaves qui se séparent au niveau 3 en « est » et « ouest ». Elles ont le rang 3. Le tchèque et le polonais sont des langues slaves « ouest » et se séparent au niveau 4. Elles ont rang 4. Enfin, le tchèque et le slovaque ont le rang 5. Les distances entre les langues sont inversement proportionnelles à leur rang : 0,2 entre le tchèque et le slovaque ; 0,4 entre le tchèque et le polonais ; 0,6 entre le tchèque et le russe ; 0,8 entre le tchèque et l'italien et 1,0 entre le tchèque et le

⁶ Voir aussi le site <http://www.ethnologue.com>

⁷ Il faut cependant noter que ce nombre est arbitraire. Il dépend de ce que l'on compte comme étant une langue à part entière. Ainsi, les Québécois parlent-ils français? On peut en douter en regardant les séries québécoises, qui lorsqu'elles passent sur les chaînes de télévision françaises sont parfois sous-titrées. Deux langues ou une langue ?

hongrois. On voit immédiatement qu'apparaît un certain arbitraire⁸, qui s'amenuisera dans ce qui suit.

Tableau 1. Arbre linguistique simplifié des langues indo-européennes et ouraliennes



Source: *Ethnologue* <http://www.ethnologue.com/subgroups/indo-european> et http://archive.ethnologue.com/15/show_family.asp?subid=91831

Le texte décrit le calcul des distances entre les langues en caractères gras.

⁸ Notons que ces distances peuvent aussi faire l'objet de pondérations. Voir Fearon (2003).

Les méthodes lexicostatistiques

Les *méthodes lexicostatistiques* sont basées sur une seule caractéristique (ou dimension) : les similarités des racines communes des mots (le lexique) dans plusieurs langues dont on peut penser qu'elles ont un ancêtre commun (par exemple l'espagnol, le français, l'italien, le portugais et le roumain). Ceci ne va pas sans problèmes. En effet, il peut exister dans ces différentes langues (ou dans des sous-groupes) : (a) des mots qui sont semblables par accident, par exemple les onomatopées ; (b) des mots qui ont été empruntés à une autre langue du groupe voire à l'extérieur du groupe⁹ et enfin, (c) des mots qui descendent d'une langue commune plus ancienne¹⁰.

Les distances lexicales sont construites à partir de mots apparentés (*cognate* en anglais), ignorant non seulement les similarités accidentelles et les mots empruntés, mais également la grammaire, la syntaxe et les autres subtilités d'une langue. Comme il serait ardu (et sans doute inutile) de comparer les lexiques formés par l'ensemble des mots dans plusieurs langues, les linguistes se sont résolus à choisir soigneusement une courte liste de mots, ou plus exactement, de « significations ». Aujourd'hui, les linguistes utilisent une liste de 200 (voire 100) significations qui sont supposées exister dans toutes les langues et cultures : animal, chasser, enfant, manger, mordre, noir, les nombres un à cinq, etc. La liste utilisée aujourd'hui résulte de légères modifications d'une liste établie par Swadesh (1952).

Voici les étapes qui permettent de calculer les distances entre les langues sur base d'une liste de significations : (a) collecter pour chaque signification de la liste les mots utilisés dans chacune des langues à étudier ; (b) établir pour chaque paire de langues la parenté des deux mots de chaque signification (oui, non, indécidable) ; (c) calculer pour chaque paire de langues le rapport « mots non-apparentés » sur « mots apparentés plus mots non-apparentés » (les indécidables sont éliminés) ; un rapport élevé (faible) indique que la distance (comprise entre 0 et 1) est importante (ou faible) ; (d) utiliser un algorithme d'analyse des données (du type *clustering*) pour partitionner les données et fabriquer un arbre.

Un exemple simple est illustré dans le Tableau 2 où nous suggérons comment on peut prendre des décisions de parenté entre 15 langues à partir de cinq significations : les nombres un à cinq. Pour faire plus facilement percevoir les proximités entre les langues, nous les avons classées en sous-groupes, de façon à

⁹ L'anglais par exemple contient 30% de mots français, suite à la conquête normande des îles britanniques. Et aujourd'hui, les Français se plaignent de l'emprunt à l'anglais de bien des mots.

¹⁰ Dyen et al. (1992) donnent l'exemple très poétique du mot *flower* emprunté au français *fleur*, alors que, de façon surprenante, ce sont *blossom* (fleurir) et *fleur* descendent du même mot ancestral.

faire apparaître plus facilement les apparentements¹¹. On voit qu'il y a des apparentements entre les quatre premiers groupes de langues, qui sont évidentes pour les nombres *deux* et *trois*, mais un peu plus subtiles (pour les non linguistes, en tout cas) dans les autres cas. Et les apparentements sont évidents, puisque toutes ces langues font partie du groupe indo-européen. Au sein de chaque sous-groupe, par exemple celui des langues germaniques, on trouve un plus grand nombre de mots apparentés, et il en est de même pour le sous-groupe constitué par les quatre langues romanes, les deux langues celtiques et les deux langues slaves, encore que même là, les dénominations des nombres *un*, *deux* et *trois* sont proches. Les deux dernières langues (hongrois et swahili) n'ont évidemment rien de commun entre elles, ni avec les autres langues. Aucune ne fait partie de la famille des langues indo-européennes. Le hongrois est arrivé en Europe centrale du temps des invasions provenant d'Asie (et fait partie des langues de l'Oural) et le swahili, parlé sur la côte orientale de l'Afrique, est d'origine bantoue, tout en ayant subi l'influence des commerçants arabes, portugais et indiens.

Tableau 2. Mots pour les nombres un à cinq dans plusieurs langues

	1	2	3	4	5
Allemand	eins	zwei	drei	vier	fünf
Anglais	one	two	three	four	five
Danois	en	to	tre	fire	fem
Néerlandais	een	twee	drie	vier	vijf
Suédois	en	två	tre	fyra	fem
Espagnol	uno	dos	tres	cuatro	cinco
Français	un	deux	trois	quatre	cinq
Italien	uno	due	tre	quattro	cinque
Portugais	um	dois	três	quatro	cinco
Breton	unan	daou (m)	tri (m)	pevar (m)	pemp
Gallois	un	dau (m)	tri (m)	pedwar (m)	pump
Polonais	jeden	dwa	trzy	cztery	piec
Russe	odfin	dva	tri	chetyre	piat'
Hongrois	egy	kettö	harom	négy	ött
Swahili	moja	mbili	tatu	nne	tano

(m) = masculin

¹¹ Il est bien évident que lors des recherches sur les apparentements, le processus termine (et ne commence pas, comme dans le Tableau 2), par les groupements.

Le Tableau 3 reproduit les distances entre quelques langues indo-européennes. On constate que les cinq langues germaniques (dans les cinq premières lignes du tableau) sont bien plus proches l'une de l'autre qu'elles ne le sont des langues romanes et du polonais. Il en est de même au sein du groupe des langues romanes, ainsi que du polonais et du russe (coin inférieur droit).

Tableau 3. Distances entre quelques langues indo-européennes (x1.000)

	Allemand	Anglais	Espagnol	Français	Italien	Polonais
Allemand	0	422	747	756	735	781
Anglais	422	0	760	764	753	761
Danois	293	407	750	759	737	749
Néerlandais	162	392	742	756	740	769
Suédois	305	411	747	756	741	763
Espagnol	747	760	0	291	212	772
Français	756	764	291	0	197	781
Italien	735	753	212	197	0	764
Portugais	753	760	126	291	227	776
Polonais	754	761	772	781	764	0
Russe	755	758	769	778	761	266

Source: Dyen, Kruskal and Black (1992), qui donnent un tableau complet de plus de 90 langues indo-européennes

Plutôt que de demander à des linguistes de décider de l'apparement des mots, Levenshtein (1966) propose d'utiliser un algorithme qui permet de comparer les mots de façon automatisée. L'idée de l'algorithme de Levenshtein est de convertir le mot d'une langue dans l'autre en insérant, en supprimant ou en substituant chaque lettre du mot de la première langue pour aboutir au mot de la deuxième (les mots sont souvent transcrits sous forme phonétique de façon à reproduire la prononciation de chaque consonne, voyelle ou phonème qui change très souvent d'une langue à l'autre). Le nombre minimum d'insertions, de suppressions ou de substitutions divisé par le nombre de lettres des deux mots représente la *distance de Levenshtein* entre les mots de même signification dans les deux langues. La distance entre les deux langues résulte d'une moyenne arithmétique des distances entre les couples de mots. Un algorithme d'analyse des données permettra, comme plus haut, de calculer un arbre.

D'autres méthodes de ce type existent et sont utilisées¹², mais tout en donnant des indications précieuses, elles s'intéressent surtout à l'histoire des langues (linguistique historique) et pas suffisamment à leur intercompréhension dans le monde dans

¹² Pour plus de détails, voir Ginsburgh et Weber (à paraître) et bien entendu McMahon et McMahon (2005).

lequel nous vivons. De plus, les distances sont symétriques : la distance entre le portugais et l'espagnol est identique à celle qui existe entre l'espagnol et le portugais¹³ et comme nous l'avons fait remarquer, les mots empruntés à d'autres langues sont exclus du calcul des distances, mais sont, dans la pratique, bien utiles.

Les méthodes mixtes : le projet ASJP

Un groupe de linguistes réunis dans l'*Automated Similarity Judgment Program* (ASJP) a entrepris de calculer les distances entre langues en combinant la lexicostatistique et quelque 85 structures phonologiques, grammaticales et lexicales. Voir Dryer et Haspelmath (2013).

Une autre approche : le temps d'apprentissage d'une langue

Chiswick et Miller (2007) proposent d'exploiter les résultats d'une étude de Hart-Gonzales et Lindemann (1993). Ceux-ci ont utilisé un échantillon d'Américains qui ont appris diverses langues (en fait 43 langues, depuis le japonais et le coréen, réputés difficiles à apprendre pour un anglophone, à l'afrikaans, au norvégien, et au suédois, qui sont comme l'anglais, des langues germaniques, donc relativement faciles à assimiler). Hart-Gonzales et Lindemann ont observé le temps qu'il a fallu à ces étudiants pour « connaître » la langue qu'ils apprenaient à différents moments de l'apprentissage. Il est évident que cette méthode est meilleure que les méthodes historiques, basées sur des approximations (nombre peu élevé de mots utilisés dans la comparaison entre deux langues, absence de prise en compte de certaines caractéristiques importantes, comme la structure des langues examinées, et les mots empruntés, etc.). Dans le cas présent, tout est pris en compte, mais il faut réaliser la difficulté qu'il y aurait à calculer de cette façon les distances entre ne fût-ce que les cent langues les plus utilisées dans le monde.

Les méthodes de calcul des distances entre langues ont toutes été utilisées à tour de rôle dans les études économiques¹⁴. Et comme nous le verrons dans la suite, l'influence sur les phénomènes socio-économiques des distances, même si elles ne sont qu'« historiques », est loin d'être négligeable.

¹³ Ce qui n'est pas le cas. Un portugais apprend plus facilement l'espagnol que l'inverse, du fait du plus grand nombre de diphtongues présentes dans la langue portugaise.

¹⁴ Ginsburgh et Noury (2008), Ginsburgh et al. (2005, 2007, 2011) utilisent les distances lexicostatistiques, Ginsburgh et al. (2014) et Melitz et Toubal (2014) ont utilisé les distances ASJP; Desmet et al. (2009, 2012, 2015) construisent leurs distances à partir d'arbres linguistiques et Chiswick et Miller (2007) ont utilisé les distances basées sur le temps d'apprentissage.

3. Distances et diversité linguistique

Considérations générales

La distance entre deux langues ne doit pas être confondue avec la distance linguistique *moyenne* entre groupes de populations dans un pays (par exemple la Suisse ou la Belgique) ou une région (par exemple l'Union Européenne), sauf si toute la population parle une langue unique. Une distance possible prend la valeur 1 s'il n'existe aucune langue commune entre les deux populations et 0 si la grande majorité de la population parle (à peu près, et nous reviendrons sur cette qualification) la même langue, comme les Allemands et les Autrichiens ou les Anglais et les Australiens. Mais une difficulté surgit lorsqu'il faut mesurer la distance linguistique entre Français et Belges ou entre Hollandais et Belges, puisque 40 pour cent des Belges parlent le français, et 60% le néerlandais.

L'idée vient alors d'estimer la probabilité que deux citoyens choisis au hasard dans les deux populations parlent la même langue. Cette probabilité sera proche de 1 entre Allemands et Autrichiens (ou Anglais et Australiens), mais sera égale à 0,4 (= 1 x 0,40) entre Français et Belges francophones qui forment 40% de la population belge et 0,6 entre Hollandais et Belges néerlandophones qui comptent pour 60% en Belgique. Les choses deviennent un peu plus compliquées si en Allemagne, 25 pourcent de la population parle le français et que 20 pourcent des citoyens français parlent l'allemand. Dans ce cas, deux citoyens pris au hasard l'un en Allemagne, l'autre en France, seront incapables de communiquer s'ils sont tout deux unilingues. La probabilité de cet événement est égale à 0,75 x 0,80 = 0,60. Le complément à 1, qui est égal à 0,40 est la probabilité que les deux individus qui se rencontrent parviennent à communiquer soit en allemand ou en français. La situation se complexifie un peu si certains Allemands et certains Français parlent également l'anglais.

De façon générale, on peut montrer que si une société est constituée de K groupes linguistiques distincts, où s_1, s_2, \dots, s_K (tous compris entre 0 et 1) représentent la dimension des groupes (en pourcent), avec $s_1 + s_2 + \dots + s_K = 1$ et d_{kl} représente la distance linguistique entre les groupes k et l , alors l'indice de diversité B (proposé en 1956 par Greenberg, un célèbre linguiste de Stanford) qui propose de mesurer la distance linguistique *moyenne* entre couples d'individus choisis au hasard dans la société ; cet indice prend la forme suivante :

$$B = \sum_{k=1}^K \sum_{l=1}^K s_k s_l d_{kl}.$$

Il convient de noter que dans le cas particulier où les distances linguistiques valent 0, si les deux groupes k et l parlent la même langue et 1 sinon, cette formule se simplifie à l'indice de diversité A (également défini par Greenberg, 1956, mais initialement proposé par Gini, 1912/1955) :

$$A = \sum_{k=1}^K \sum_{l=1}^K s_k s_l,$$

pour tout couple $k \neq l$, qui, étant donné que $(\sum_{k=1}^K s_k)^2 = 1$, est souvent noté $1 - \sum_{k=1}^K s_k^2$.

Ces deux types d'indices (mais plus particulièrement l'indice A) ont été souvent utilisés dans des équations économétriques du type :

$$y = \theta IND + \sum_h \xi_h z_h + \eta \quad (*)$$

où y est une variable indépendante (une mesure du développement économique, la corruption, les biens publics, ...) dont on pense qu'elle peut être influencée par la diversité ethnique, ou linguistique résumée par IND qui représente l'indice A ou B , les z_h sont des variables de contrôle et η est un terme d'erreur. Les paramètres θ et les ξ_h doivent être estimés ; θ est le paramètre d'intérêt qui mesure l'influence de la diversité sur la variable du membre de gauche¹⁵.

Pour construire les deux indices, il faut définir les groupes $k = 1, 2, \dots, K$. Quand faut-il distinguer deux groupes linguistiques, quand ne faut-il pas le faire ? L'italien est-il suffisamment différent du vénitien pour décider d'en faire deux groupes différents ? Comment traiter le sicilien ? Considérons par exemple les cas de la Principauté d'Andorre et de la Belgique¹⁶. Andorre est formé de deux groupes d'environ 50 pourcent ; un des groupes parle espagnol, l'autre le catalan, menant à un indice A égal à 0,50. En Belgique, comme nous l'avons écrit plus haut, il y a 40 pourcent de francophones et 60 pourcent de néerlandophones, ce qui donne un indice A égal à 0,48. La diversité en Andorre et en Belgique est par conséquent pratiquement la même. A ceci près que la distance lexicographique entre les deux langues en Andorre est 0,15, alors qu'elle est de 0,76 en Belgique. L'indice B mesure donc bien mieux la diversité « réelle » qui est évidemment plus élevée en Belgique qu'en Andorre. Remarquons cependant que l'on peut considérer d'autres mesure de diversité que la diversité linguistique : la diversité ethnique en est une, la diversité religieuse en est

¹⁵ Voir Alesina et al. (1999), Alesina et al. (2003), Alesina et al. (2004), Alesina et La Ferrara (2005), Alesina et Zhuravskjaya (2011).

¹⁶ Les exemples sont basés sur Desmet et al. (2009).

une autre, mais les indices tels qu'ils existent aujourd'hui ne permettent pas de tenir compte de plus d'un type de diversité à la fois.

Comment constituer les groupes ?

La question qu'il faut se poser est donc double :

(a) sous quel angle faut-il considérer la diversité d'une population lorsqu'on veut étudier son effet sur les comportements économiques ? Nous renvoyons à Alesina et al. (2003) pour un essai de réponse à la question et nous nous intéresserons dans la suite uniquement à des réponses à la question suivante :

(b) quel est le degré de diversité linguistique (ou autre) qu'il convient de considérer. Faut-il ou non considérer le vénitien comme étant différent du sicilien et séparer les populations de l'Italie en plusieurs groupes ?

Desmet et al. (2009) ont été parmi les premiers à montrer que l'introduction d'un indice de type *B* qui tient compte des distances linguistiques offre de meilleurs résultats qu'un indice de type *A*. Dans leur article, les auteurs examinent l'impact des deux indices sur la redistribution des revenus (représentée par les transferts et subventions en pourcentage du PIB) dans plus d'une centaine de pays. Ils constatent que l'incorporation des distances linguistiques (calculées à partir des arbres linguistiques d'*Ethnologue*) rend l'effet de la diversité significatif, aussi bien sur le plan statistique que sur le plan économique, alors que l'indice *A* conduit à des résultats non significatifs.

Forts de cette conclusion encourageante, Desmet et al. (2012, 2015) ont étudié l'impact de différents types de clivages (ou agrégations) des groupes linguistiques sur l'économie : au niveau le plus élevé (ou grossier) d'agrégation¹⁷, seules les grandes familles linguistiques telles que l'ensemble des langues indo-européennes ou l'ensemble des langues sino-tibétaines qui se sont séparées de la racine il y a plusieurs milliers d'années sont importantes. De tels clivages génèrent des indices de type *B* qui permettent une meilleure explication de certains phénomènes que d'autres. Les séparations anciennes sont davantage reliées à la nature profonde des populations, à leur culture, à leur ethnicité, aux attitudes et aux relations de confiance que les séparations plus récentes telles que celles représentées par les différences entre l'italien et le vénitien.

Ce que font par la suite Desmet et al. (2012, et à paraître), est d'estimer les paramètres de la relation (*) pour des niveaux d'agrégation linguistiques différents. Ils constatent que la redistribution des revenus est mieux expliquée par des agrégations plus

¹⁷ Voir <http://www.ethnologue.com/browse/families>

grossières, alors que des agrégations plus fines sont plus aptes à expliquer les différences de croissance économique.

4. Distances linguistiques et « comportements » économiques

L'analyse des applications économiques est guidée par une structure commune inspirée par la physique newtonienne dans laquelle la force d'attraction entre deux corps est directement proportionnelle au produit des masses des deux corps et inversement proportionnelle à (en fait au carré de) la distance qui les sépare.

Les économistes, à commencer par Tinbergen (1962), se sont approprié cette formulation pour expliquer les choix des agents économiques relatifs aux flux de commerce international, aux flux migratoires entre pays i et j , aux flux de traductions littéraires entre langues i et j , ainsi qu'à l'apprentissage d'une langue j par des locuteurs de la langue i . Dans la suite, nous décrirons ces différentes applications, mais limiterons la discussion à la mesure et à l'impact de la distance linguistique, qui a dans tous les cas l'effet négatif attendu. Ainsi, une distance plus importante entre langues utilisées réduit les flux commerciaux, les flux migratoires, les flux de traduction, le nombre d'individus nés dans une langue qui acquièrent une langue étrangère et influence le comportement de ceux qui élisent les vainqueurs au concours Eurovision de la chanson : on votera davantage pour un concurrent dont la langue est proche.

L'équation type qui sera estimée dans les différentes situations peut s'écrire :

$$y_{ij} = \alpha_0 + \alpha_1 m_i + \alpha_2 m_j + \beta d_{ij} + \sum_k \gamma_k z_{k,ij} + \varepsilon_{ij},$$

où y_{ij} est un flux (par exemple, les exportations du pays i vers le pays j), m_i et m_j sont des variables représentatives des pays i et j (leurs PIBs par exemple), d_{ij} est la distance linguistique entre i et j , les $z_{k,ij}$ d'autres variables de contrôle (taux d'alphabetisation des deux pays, qualité de la chanson, etc.), et les α, β, γ_k des paramètres qu'il faut estimer; enfin ε_{ij} est une erreur non observable, jouissant des propriétés statistiques habituelles. Les signes attendus des principaux paramètres sont identiques à ceux de la mécanique newtonienne : α_1 et $\alpha_2 > 0$ (le flux est d'autant plus important que les « masses » m_i et m_j le sont) et $\beta < 0$: la distance, qui est la variable d'intérêt, dans notre cas, exerce une influence négative sur les flux).

Commerce international

L'article de Tinbergen a été suivi par un très grand nombre d'articles sur la question¹⁸. Pour rester simple et aller à l'essentiel, dans la plupart de ces articles, le flux commercial entre deux pays est fonction de variables qui décrivent les deux pays (par exemple leurs produits nationaux bruts, ou tout simplement des effets fixes de source et de destination des flux), de variables de contrôle telles que leur proximité géographique, leur éventuel statut d'ancienne colonie, etc. et bien entendu les distances linguistiques entre pays, puisque leurs agents commerciaux doivent être capables de communiquer et de signer les contrats nécessaires. Dans les premiers travaux, les chercheurs se sont satisfaits d'une distance qui prenait deux valeurs : 0 si la langue entre les partenaires était commune et 1 autrement. Les mesures se sont diversifiées et affinées. Nous décrivons ici l'étude récente de Melitz et Toubal (2014) qui généralise l'approche considérée par Melitz (2008).

Au lieu d'une seule distance entre la langue du pays exportateur et celle du pays importateur, Melitz et Toubal (2014) considèrent quatre distances différentes qui ont des rôles différents :

(a) s'il existe une langue maternelle commune entre les deux pays, la variable *CNL* (common native language) prend pour valeur la probabilité que deux habitants choisis au hasard — un dans chaque pays — la partagent ; cette probabilité est égale au produit des parts des habitants qui parlent cette langue dans chaque pays. Si tous les habitants la parlent, cette probabilité vaut 1. S'il n'y a pas de langue commune, la probabilité est 0.

(b) s'il existe une autre langue commune, une variable *CSL* (common spoken language) est construite de la même façon que la première, mais avec d'autres parts de population ;¹⁹

(c) s'il existe une langue officielle commune, la variable *COL* (common official language) est égale à 1, sinon elle est égale à 0;

(d) dans tous les cas, la distance linguistique *LP* (linguistic proximity) entre langues maternelles du pays exportateur et du pays importateur.²⁰

¹⁸ Voir notamment Anderson et van Wincoop (2004), Egger et Lassaïn (2012) pour une revue de la littérature et Anderson (1979) pour les fondements théoriques de l'approche gravitationnelle proposée par Tinbergen, restée jusque-là purement empirique. Voir aussi bien entendu Melitz (2015) et Carrère et Masood (2015) dans ce livre.

¹⁹ Une formulation similaire a été proposée en 2009 par Fidrmuc and Fidrmuc (2009).

²⁰ On remarquera que contrairement aux autres cas décrits où ce sont des distances entre paires de langues qui sont utilisées, dans *CNL* et *LP*, Melitz et Toubal utilisent des distances moyennes décrites dans la partie 3.

Les raisons de ces quatre choix sont les suivantes. Si l'estimation des paramètres du modèle indique que le paramètre estimé pour *CSL* est significativement différent de zéro (au sens statistique) — nous dirons plus simplement « est significatif » — en la présence de *CNL*, cela implique que la facilité de communication est basée sur plus que la seule ethnicité commune et la confiance engendrée par une langue maternelle commune (*CNL*). La « significativité » de *COL* en présence de *CNL* et *CSL*, est une indication de l'importance du support institutionnel qui permet la traduction d'une langue vers d'autres. La « significativité » de *LP* en présence des trois autres variables, suggère qu'il n'est pas trop difficile ni trop coûteux de trouver des traducteurs ou des interprètes lorsque les langues maternelles diffèrent.

Les résultats de Melitz et Toubal montrent que l'utilisation simultanée des quatre distances, qui toutes ont un effet significatif (et une raison d'être différente) sur les flux commerciaux internationaux améliore sensiblement la prise en compte du rôle des langues pour expliquer les échanges (et inversement, ce qui cause un problème de biais dans ce cas-ci). La pratique habituelle des économistes d'utiliser une seule variable qui peut prendre deux valeurs (0 ou 1) pour la présence ou l'absence d'une langue commune est largement insuffisante.

Migrations

L'approche usuelle pour analyser la décision d'émigrer est basée sur les coûts et bénéfices que cette migration apportera à l'individu. La perspective d'une rémunération plus élevée est comparée aux coûts monétaires et psychologiques, à l'ajustement à de nouvelles conditions de vie et à l'éventuel déracinement d'une partie de la famille si elle accompagne le migrant et à son isolement si elle reste dans le pays natal.

La forme de l'équation estimée pour évaluer les facteurs qui motivent la décision est similaire à l'équation gravitationnelle dont il a été question plus haut, mais les fondements en sont différents. Massey et al. (1993) et Carrington et al. (1996) suggèrent que les réseaux de migrants dans le pays de destination créent les externalités positives qui permettent d'attirer les migrants, en rendant leur assimilation plus aisée.

Un exemple récent et représentatif est l'application proposée par Beine et al. (2011) qui étudient les flux migratoires entre 1990 et 2000 de 195 pays vers 30 pays de l'OCDE, en distinguant migrants qualifiés et peu qualifiés. La variable qui exerce la plus grande influence sur les migrants est la taille de la diaspora dans le pays d'immigration, mais d'autres incitants, tels que le salaire et le degré de générosité du pays receveur (notamment en termes de sécurité sociale et d'accès aux soins), une langue commune entre pays d'émigration et d'immigration (variable binaire) joue

également un rôle important, surtout lorsque les migrants sont qualifiés. Ils retrouveront un emploi d'autant plus facilement qu'ils connaissent la langue du pays vers lequel ils migrent.

Transmissions culturelles et traductions littéraires

Les recherches sur la transmission culturelle traitent essentiellement des media, et plus spécialement du cinéma et des programmes de télévision²¹, qui nécessitent traduction, doublage ou sous-titres. Bien que la radio et la télévision aient considérablement changé le mode de transmission de la « culture », l'écrit demeure essentiel. Ce n'est pas tant l'écrit scientifique qui est concerné ici mais l'écrit littéraire. Si l'anglais est devenu la langue scientifique par excellence aujourd'hui, et un grand nombre de chercheurs rédigent directement le résultat de leurs recherches dans cette langue, ce n'est évidemment pas le cas des écrits littéraires qui, dans leur grande majorité sont difficiles à comprendre et même à « sentir » s'ils ne sont pas traduits dans la langue de celui qui les lit. Mais ici encore, l'anglais est accusé de domination, par les sociologues, et par certains économistes comme Melitz (2007). Ganne et Minon (1992) ont été parmi les premiers à faire remarquer que l'Allemagne, l'Espagne, la France, l'Italie traduisent bien plus (respectivement 15, 26, 18 et 25 pourcent de la production domestique) que la Grande Bretagne (3,3 pourcent seulement). Dans ce cas particulier, on pourrait (et c'est ce que font Ganne et Minon) attribuer le faible nombre de traductions en Grande Bretagne au fait qu'un grand nombre d'ouvrages qui y sont vendus proviennent des Etats-Unis. Un peu plus tard, Heilbron (1999) fera remarquer que 50 à 70 pourcent des traductions dans les pays européens sont faites à partir de l'anglais²².

Ginsburgh et al. (2011) (GGW dans la suite) notent néanmoins que la population dont la langue maternelle est l'anglais (quelque 360 millions) est bien plus importante que celle des autres langues européennes. En outre, l'anglais est parlé (et écrit) dans un grand nombre de cultures très différentes (les Etats-Unis, le Canada, l'Inde, l'Afrique de l'est et de l'ouest, l'Australie, etc.), ce qui donne à leurs romans un aspect beaucoup plus varié que ne peut le faire l'allemand par exemple. GGW mettent aussi en évidence l'importance des proximités culturelles. Un roman policier qui se déroule à New York sera sans doute plus rapidement traduit en français qu'un roman du même genre qui se passe en Chine ou en Estonie. Pensez aussi à la difficulté à lire Dostoïevski ou Tolstoï, dont le nombre de personnages (pour lesquels il existe souvent une liste en début de volume) et leurs noms nous sont peu familiers, comme le sont d'ailleurs les noms des rues à Shanghai ou à Talinn où se poursuivent les inspecteurs et les bandits. Pour vérifier cette intuition, GGW construisent un modèle

²¹ Voir par exemple Hoskins et al. (1997).

²² Voir Heilbron and Sapiro (à paraître) pour un aperçu beaucoup plus complet sur l'approche sociologique des processus de traduction.

dans lequel l'équation de demande d'ouvrages traduits de la langue i vers la langue j est fonction des populations qui parlent les deux langues (en faisant l'approximation que le nombre de livres écrits et lus est proportionnel aux populations) et inversement proportionnel à la distance culturelle (représentée ici, faute de mieux, par la distance lexicostatistique entre les deux langues). Et nous retrouvons une fois encore le modèle gravitationnel, dans lequel la distance joue un rôle important : une augmentation de 10 pourcent de la distance entre les langues réduit le nombre de traductions de l'une à l'autre de 10 pourcent environ²³.

Apprentissage des langues étrangères

Selten et Pool (1991) sont les premiers à avoir introduit un modèle de communication basé sur la théorie des jeux qui prend en compte les coûts et les avantages auxquels ceux qui apprennent une langue étrangère peuvent s'attendre. Ils montrent que ce jeu possède un équilibre (ou une solution). Church et King (1993) simplifient ce modèle en considérant deux langues et deux populations (deux pays). Dans ce modèle, le bénéfice communicationnel de l'individu d'un pays (qui connaît sa langue maternelle) s'accroît avec le nombre de ceux avec lesquels il peut communiquer dans sa propre langue et/ou dans l'autre. Chaque individu peut calculer son coût (commun à tous les individus de sa population) et son bénéfice et décider (ou non) d'apprendre l'autre langue. Etant donné que les coûts d'apprentissage sont identiques pour tous les individus d'une population, il n'existe pas de solution à ce jeu dans laquelle chacune des deux populations apprend la langue de l'autre²⁴. Gabszewicz et al. (2011) introduisent des aptitudes hétérogènes à apprendre l'autre langue, ce qui leur permet de montrer qu'il existe des solutions dans lesquelles une partie de chacune des deux populations apprend la langue de l'autre, ce qui a amené Ginsburgh et al. (2007), et, plus tard, Ginsburgh et al. (2014) à estimer une équation dans laquelle les décisions des habitants de quelque 190 pays à apprendre une autre langue que leur langue maternelle, c'est-à-dire une des 13 langues les plus parlées dans le monde est fonction d'un certain nombre de variables, dont le nombre de locuteurs de la langue maternelle, celui de la langue acquise et la distance (ASJP) entre les deux langues, une formulation à nouveau proche du modèle gravitationnel. Une augmentation de 10 pourcent de la distance entre langue maternelle et langue acquise réduit de 6 pourcent la probabilité de se mettre à l'étude de la langue étrangère.

Le concours Eurovision de la chanson

Les résultats du concours Eurovision de la chanson sont souvent l'objet de controverses qui mettent l'accent sur l'aspect politique des votes : les pays forment

²³ Dans la foulée du modèle, GWW montrent également que, compte tenu des populations, l'anglais n'est pas la langue la plus traduite (par tête).

²⁴ En fait, soit toute la population d'un pays apprend l'autre langue, soit personne ne l'apprend et c'est la population de l'autre pays qui apprend, soit aucune des deux populations ne bouge.

des cliques dont les membres se coordonnent (implicitement ou explicitement) pour voter les uns pour les autres. Ginsburgh et Noury (2008) ont analysé les résultats des concours annuels entre 1975 et 2003 en estimant une équation de la forme gravitationnelle dans laquelle le nombre de votes émis dans le pays *i* en faveur du candidat du pays *j* (il n’y a qu’un seul candidat par pays) est fonction d’un certain nombre de variables décrivant les pays, de la qualité estimée de la chanson, de la distance linguistique lexicographique entre électeurs et candidats. Ils montrent que dès que la distance linguistique est prise en compte, les affinités politiques supposées disparaissent. Ceci ne signifie pas qu’il n’y a pas un certain biais dans la manière de voter, mais il est expliqué par la proximité linguistique qui peut être qualifiée de culturelle.

5. Exclusion et inclusion linguistique

La gestion des langues dans un pays où coexistent plusieurs langues dites « officielles » (trois en Suisse comme en Belgique, une seule en France) pose des problèmes aussi bien sur l’entente entre populations que sur la gestion elle-même puisque tous les documents officiels ou légaux doivent exister dans toutes les langues. Le problème se pose de façon aiguë dans l’Union Européenne (UE) qui doit maîtriser 24 langues à un coût annuel qui dépasse probablement assez largement un milliard d’euros. Même si dans le cas de l’UE elle reste purement théorique, puisque tout changement du régime linguistique actuel nécessiterait l’unanimité des 28 Etats-membres, on peut se poser la question de ce qui se passerait si on réduisait le nombre de langues officielles. Cette question s’est d’ailleurs posée bien souvent dans le monde, y compris en France (depuis l’édit de Villers-Cotterêts pris en 1539 par François I), en Espagne (décrets *Nueva Planta* pris par Philippe V entre 1707 et 1716), en Grande Bretagne (*Act of Union* pris en 1536 par Henry VIII), ou en Russie (sous le tsar Pierre le Grand surtout), lorsque les rois, princes ou tsars ont voulu unifier leur pays. Et elle continue de se poser avec grande acuité dans certains pays asiatiques ou africains, comme le Nigéria, où coexistent 522 langues, même si quatre langues (anglais, hausa, igbo et yoruba) sont plus pratiquées que les autres.

Ginsburgh et Weber (2005) et Ginsburgh et al. (2005) ont suggéré de procéder à des simulations du nombre de locuteurs qui se verraient privés de leurs droits linguistiques (*disenfranchised*) dans l’UE si certaines des langues perdaient leur statut de langue « officielle ». Il est possible de faire ces calculs à partir de recensements ou d’enquêtes qui demandent à chaque citoyen de donner une liste des langues qu’il est capable de parler et/ou de comprendre avec une « certaine » aisance²⁵. Il suffit alors

²⁵ On peut évidemment contester le bien fondé des réponses qui sont données et qui pèchent sans doute souvent par exagération.

d'« empiler » les langues que l'on pense devoir conserver et de mesurer le nombre de citoyens dont l'application réelle de cette mesure ferait des « exclus linguistiques ».

Toute simple qu'apparaisse pareille procédure, elle pose néanmoins un problème et mérite une extension. Le problème est qu'il faut exclure du décompte final ceux qui parlent simultanément deux ou plus des langues retenues dans la simulation, sans quoi on les compterait deux ou plusieurs fois. Ainsi si les langues retenues sont par exemple le français, l'anglais et l'allemand, il faut soustraire ceux qui parlent deux des trois voire, les trois langues. L'extension, quant à elle, consiste à tenir compte des distances linguistiques entre les langues retenues et les langues qui ne le sont pas. En effet, à supposer que l'on ôte de la liste des langues officielles le portugais, mais que l'on retienne l'espagnol, le locuteur portugais pourrait se sentir moins exclu si l'espagnol fait partie des langues retenues, étant donné la proximité des deux langues.

La procédure de sélection des langues retenues est incrémentiel : les langues deviennent « officielles » une à une, et chaque langue est introduite dans l'ensemble des langues officielles de façon à minimiser le taux d'exclusion global au niveau de l'UE²⁶.

Le Tableau 4 donne une vue d'ensemble de la situation. Les deux premières colonnes donnent le nom du pays et de sa population, avec, en dernière ligne, l'ensemble de l'UE à 28 pays. Le reste du tableau est divisé en trois parties distinctes, renfermant chacune le pourcentage de la population (dans un pays ou dans l'ensemble de l'UE) qui serait partiellement ou totalement privée de ses droits linguistiques et en tout cas mise à mal dans sa compréhension des documents légaux et des discussions au Parlement Européen dans une situation linguistique (simulée) donnée.

Dans les six premières colonnes dont le titre général est « langues individuelles », les pourcentages sont donnés par langue individuelle (GB pour l'anglais, D pour l'allemand, F pour le français, I pour l'italien SP pour l'espagnol et PL pour le polonais, les six langues les plus parlées dans l'UE). Ainsi, dans la ligne « Allemagne » et la colonne « GB » on trouve le pourcentage de la population allemande qui dit ne comprendre l'anglais (62%). Par contre si, comme dans la colonne « D », l'allemand devenait la seule langue officielle, 1% seulement de la population allemande (résultant de l'immigration) ne comprendrait pas, etc. Dans la dernière ligne, on trouve les pourcentages pour l'ensemble de l'UE. Ces pourcentages indiquent que si l'anglais était appelé à devenir la seule langue officielle, 62,7% de la population de l'UE serait « exclue » (ou encore, que 37,3% de la population a déclaré comprendre l'anglais). Il faut noter que si l'une des cinq autres langues était choisie, la situation serait pire encore, puisque les pourcentages deviennent de plus en plus élevés :

²⁶ Voir Fidrmuc et al. (2007) pour des explications plus détaillées de la procédure utilisée.

75,2% pour l'allemand, 80,3% pour le français, etc. Ceci montre, entre autres, que le français est la troisième langue la plus parlée dans l'UE, mais reste néanmoins incomprise par un peu plus des quatre cinquièmes de la population européenne.

Dans les six colonnes suivantes, les langues sont introduites l'une après l'autre, dans un ordre qui minimise l'exclusion linguistique dans l'UE. Ainsi, la première langue est évidemment l'anglais, qui est le plus fréquemment utilisé (par 37,3% des européens). Vient ensuite l'allemand (ce qui est indiqué par +D). Deux langues sont à présent considérées « officielles » simultanément, l'anglais et l'allemand, ce qui nous fournit les taux d'exclusion qui figurent dans la deuxième de ces six colonnes, et le taux d'exclusion dans l'UE tombe à 49,6%. La langue suivante, toujours choisie dans le but de minimiser l'« exclusion » des citoyens dans l'UE est le français (colonne +F), qui réduit le taux d'exclusion à 38,1%. Viennent ensuite, à tour de rôle, l'italien (+I), l'espagnol (+SP) et le polonais (+PL). En dernière ligne, on voit que si les six langues (GB+D+F+I+SP+PL) sont officielles, le taux d'exclusion tombe à 16,8%. Il n'en reste pas moins que dans certains pays, tels que la Bulgarie, la Hongrie, la Lettonie, et le Portugal, la population qui ne comprend aucune de ces langues s'élève à plus de 75%. La procédure peut se poursuivre en empilant successivement d'autres langues (voir Fidrmuc et al., 2007), mais les gains obtenus par l'introduction d'une langue supplémentaire par les citoyens deviennent, on s'en doute, de plus en plus faibles.

Dans les six dernières colonnes, les langues sont à nouveau empilées ; mais, dans l'ordre d'entrée, on tient compte des distances entre ces langues. Une faible distance (comme entre l'anglais et l'allemand) permettrait donc aux allemands de « comprendre » l'anglais si ce dernier était la seule langue officielle. Et effectivement, on voit que le pourcentage des allemands qui ne comprennent pas l'anglais tombe de 62% (si l'on ne tient pas compte des distances) à 26% s'il en est tenu compte²⁷. Comment peut-on expliquer que les langues n'acquièrent pas le statut (simulé) de langue officielle dans le même ordre que lorsqu'elles sont empilées sans tenir compte des distances ?

²⁷ La distance lexicographique entre l'anglais et l'allemand est égale à 0,42 (voir Tableau 1), ce qui, intuitivement, peut être interprété comme signifiant que 58% des Allemands qui ne connaissent pas l'anglais le comprennent néanmoins facilement. Et effectivement, si l'on ne tient pas compte de la distance linguistique entre les deux langues, 62% des Allemands ne connaissent pas l'anglais (voir plus haut dans le texte et dans le Tableau 4, à l'intersection de la ligne Allemagne et de la colonne GB). Mais la faible distance entre les deux langues fait que 58% de ces 62% le comprennent sans le « connaître ». Il en résulte qu'aux 38% qui le connaissent, il faut ajouter $0,62 \times 0,58 = 36\%$ qui le « connaissent » grâce à la faible distance, ce qui nous amène à un total de $38\% + 36\% = 74\%$ qui le connaissent directement ou indirectement et que seulement $100\% - 74\% = 26\%$ ne comprennent pas (Tableau 4, ligne Allemagne, colonne GB dans la dernière partie du tableau).

Tableau 4. Locuteurs privés de compréhension
(en % de la population dans chaque pays et dans d'UE28)

	Pop.	Langues individuelles						Langues empilées (sans distances ling.)						Langues empilées (avec distances ling.)					
		GB	D	F	I	SP	PL	GB	+D	+F	+I	+SP	+PL	GB	+F	+D	+PL	+SP	+H
Austria	8.2	55	1	94	95	98	100	55	0	0	0	0	0	23	23	0	0	0	0
Belgium	10.4	59	87	29	97	97	99	59	56	18	18	18	18	33	8	3	3	3	3
Bulgaria	7.8	84	94	96	99	99	100	84	81	79	79	79	78	64	62	60	28	28	28
Croatia	4.4	71	85	99	93	99	100	71	62	62	60	60	60	51	49	46	20	20	20
Cyprus	0.7	49	98	95	99	99	100	49	49	49	48	48	48	41	40	39	39	39	39
Czech R.	10.2	84	81	98	100	100	98	84	70	69	69	69	67	59	58	51	16	16	16
Denmark	5.4	34	73	97	99	98	100	34	31	31	31	31	30	14	14	9	9	9	9
Estonia	1.3	75	92	100	100	100	100	75	70	70	70	70	69	60	60	58	34	34	28
Finland	5.2	69	95	99	100	100	100	69	67	67	67	67	67	65	65	64	64	64	45
France	60.6	80	95	1	95	93	100	80	77	1	1	0	0	60	0	0	0	0	0
Germany	82.5	62	1	92	99	98	98	62	1	1	1	1	1	26	26	1	0	0	0
Greece	11.1	68	94	95	98	100	100	68	64	63	63	63	63	55	54	51	50	50	50
Hungary	10.1	92	91	100	99	100	100	92	85	85	85	85	85	88	87	84	84	84	0
Ireland	4.1	1	98	91	100	99	99	1	1	1	1	1	1	1	1	1	1	1	1
Italy	58.5	75	96	90	3	97	100	75	74	69	1	1	1	57	15	14	14	14	14
Latvia	3.4	86	96	99	100	100	87	86	82	82	82	82	71	65	64	62	27	27	27
Lettonia	2.3	85	97	100	100	100	99	85	83	83	83	83	82	65	64	63	27	27	27
Luxemburg	0.5	61	12	11	95	99	100	61	8	1	1	1	1	28	4	0	0	0	0
Malta	0.4	32	99	95	65	99	100	32	31	31	31	31	31	31	31	31	31	31	31
Netherlands	16.3	23	43	81	100	97	100	23	18	18	18	18	18	9	9	3	3	3	3
Poland	38.2	82	90	99	99	100	2	82	77	76	76	76	1	61	60	57	1	1	1
Portugal	10.5	85	98	91	99	96	100	85	84	81	81	79	79	64	24	24	24	10	10
Romania	21.7	86	97	90	98	99	100	86	85	81	80	79	79	67	35	35	34	33	30
Slovakia	5.4	83	82	99	100	100	98	83	72	72	72	72	70	59	59	53	17	17	13
Slovenia	2.0	59	79	98	91	99	100	59	50	50	45	45	45	41	39	34	17	17	17
Spain	43.0	84	98	94	99	2	100	84	84	81	80	2	2	64	22	22	22	1	1
Sweden	9.0	33	88	97	99	99	100	33	33	33	33	33	33	14	14	10	10	10	10
Un.Kingdom	60.0	1	98	91	99	98	100	1	1	1	1	1	1	1	1	1	1	1	1
UE28	485.1	62.7	75.2	80.3	86.8	89.0	91.7	62.7	49.6	38.1	29.7	22.8	16.8	43.2	24.2	18.4	11.5	9.3	7.2

Pop = Population (en millions); GB = Anglais; D = Allemand; F = Français; I = Italien; SP = Espagnol; PL = Polonais; H = Hongrois

Tout d'abord, étant donné que l'allemand qui entrerait comme deuxième langue est proche de l'anglais (première langue entrée), son entrée serait retardée, puisque les 82,5 millions d'allemands « comprennent » maintenant l'anglais qui leur suffit, et c'est le français, qui est bien plus éloigné de l'anglais (distance lexicographique égale à 0,76) qui devient nécessaire pour réduire le taux d'incompréhension en le faisant passer de 43,2% dans l'UE à 24,2% (dernière ligne, 6 dernières colonnes du Tableau 4). L'allemand suit, puis arrive le polonais (une langue slave distante des trois premières, mais qui a l'avantage de faire tomber les taux d'exclusion linguistique en Bulgarie, Croatie, Lettonie, Lituanie, Slovaquie, République tchèque et Slovénie, pays dont les langues font toutes partie de la famille slave. L'espagnol suit et le « groupe des six » langues de l'exemple se termine par le hongrois, une langue non-indoeuropéenne, mais qui réduit évidemment le taux d'exclusion pour 10 millions de hongrois, quelques estoniens et finlandais (dont les langues font partie du même groupe que le hongrois), et quelques frontaliers slovaques et roumains.

Ces données simulées permettraient, le cas échéant, aux autorités de l'UE de prendre des décisions sur les langues qui resteraient « officielles »²⁸. Notre exemple est limité à l'espace européen pour lequel existent des données, mais rien n'empêche d'utiliser ce type de méthode pour examiner comment traiter et réduire la diversité linguistique dans les pays de l'Afrique sub-saharienne, où dans certains cas plusieurs centaines de langues coexistent et rendent souvent la cohabitation difficile (Nigéria, République du Congo) et réduisent le commerce intérieur, encore qu'au Nigéria par exemple, les populations aient finalement choisi de se coordonner sur quatre langues véhiculaires plus largement parlées.

6. Conclusions

Il importe de reconnaître que les économistes ont, depuis longtemps, essayé d'intégrer les langues comme variables explicatives des comportements des agents. L'approche était très simple dans les premières études relatives au commerce international et était limitée à la question « y a-t-il une langue commune entre les pays i et j ». Elle s'est largement complexifiée dans les études de commerce international, comme en témoigne l'article de Melitz et Toubal (2014) qui utilisent plusieurs types de distances répondant à des préoccupations différentes. Et les distances linguistiques diverses se sont introduites dans d'autres domaines de recherches, parfois imprévus comme le Concours Eurovision de la Chanson.

Références

- Alesina, A., Baqir, R., et Easterly, W. (1999), "Public goods and ethnic divisions", *Quarterly Journal of Economics* 114, 1243-1284.
- Alesina, A., Baqir, R., et Hoxby, C. (2004), "Political jurisdictions in heterogeneous communities", *Journal of Political Economy* 112, 349-396.
- Alesina, A., Devleeschouwer, A., Easterly, W., Kurlat, S., et Wacziarg, R. (2003), "Fractionalization", *Journal of Economic Growth* 8, 155-194.
- Alesina, A., et La Ferrara, E. (2005), "Ethnic diversity and economic performance", *Journal of Economic Literature* 43, 762-800.
- Alesina, A., et Zhuravskaya, E. (2011), "Segregation and the quality of government in a cross-section of countries", *American Economic Review* 101, 1872-1911.
- Anderson, J. (1979), "A theoretical foundation for the gravity equation", *American Economic Review* 69, 106-116.
- Anderson, J., et van Wincoop, E. (2004), "Trade costs", *Journal of Economic Literature* XLII, 691-751.
- *Atlas Narodov Mira* (1964), The Miklucho-Maklai, Ethnological Institute at the Department of Geodesy and Cartography of the State Geological Committee of the Soviet Union.

²⁸ Fidrmuc et al. (2009) proposent l'utilisation de procédures de vote, en supposant que l'unanimité (qui est actuellement la règle en matière linguistique) ne soit plus requise.

- Beine, M., Docquier, F. et Özgen, C. (2011), "Diasporas", *Journal of Development Economics*, 95, 30-41.
- Carrington, W., Detragiache, E., et Vishwanath, T. (1996), "Migration with endogenous moving costs", *American Economic Review* 86, 909-930.
- Cavalli-Sforza, L. L. (2000), *Genes, Peoples, and Languages*, Berkeley, CA: University of California Press.
- Chiswick, B., et Miller, P. (2007), "Linguistic distance. A quantitative measure of the distance between English and other languages", dans Chiswick, B., et Miller, P. (dir.), *The Economics of Language. International Analyses*, London and New York: Routledge.
- Church, J., et King, I. (1993), "Bilingualism and network externalities", *Canadian Journal of Economics* 26, 337-345.
- Desmet, K., Ortuno-Ortin, I., et Wacziarg, R. (2012), "The political economy of linguistic cleavages", *Journal of Development Economics* 97, 322-338.
- Desmet, K., Ortuno-Ortin, I., et Wacziarg, R. (2015), "Linguistic cleavages and economic development", dans Ginsburgh, V., et Weber, S. (dir.), *The Palgrave Handbook of Economics and Language*, Houndmills Basingstoke: Palgrave Macmillan.
- Desmet, K., Ortuno-Ortin, I., et Weber, S. (2009), "Linguistic diversity and redistribution", *Journal of the European Economic Association*, 7, 1291-1318.
- Dryer, M., et Haspelmath, M. (dir.), (2013), *The World Atlas of Language Structures Online*, Leipzig: Max Planck Institute for Evolutionary Anthropology (Available online at <http://wals.info>, accessed on December 27, 2014.)
- Dyen, I., Kruskal, J., et Black, P. (1992), "An Indoeuropean Classification, a Lexicostatistical Experiment", *Transactions of the American Philosophical Society* 82/5.
- Egger, P., et Lassmann, A. (2012), "The language effect in international trade: a meta-analysis", *Economics Letters* 116, 221-224.
- Ethnologue (2009), *Languages of the World*, Dallas, TX: SIL International.
- Fearon, J. (2003), "Ethnic and cultural diversity by country", *Journal of Economic Growth* 8, 195-222.
- Fearon, J., et Laitin, D. (1999), "Weak states, rough terrain, and large ethnic violence since 1945", Paper presented at the annual meetings of the American Political Science Association, Atlanta, GA.
- Fidrmuc, J., et Fidrmuc, J. (2009), "Foreign languages and trade", CEPR Discussion Paper 7228.
- Fidrmuc, J., Ginsburgh, V., et Weber, S. (2007), "Ever closer Union or Babylonian discord? The official-language problem in the European Union", CEPR Discussion Paper 6367.
- Fidrmuc, J., Ginsburgh, V., et Weber, S. (2009), "Voting on the choice of core languages in the European Union", *European Journal of Political Economy* 25, 56-62.

- Gabszewicz, J., Ginsburgh, V., et Weber, S. (2011), "Bilingualism and communicative benefits", *Annals of Economics and Statistics* 101/102, 271-286.
- Ganne, V., et Minon, M. (1992) « Géographies de la traduction », dans Barret-Ducrocq, F. (dir.), *Traduire l'Europe*, Paris: Payot.
- Gini, C. (1912/1955), Variabilità e mutabilità, *Studi Economico-Giuridici della R. Università di Cagliari* 3, 3-159. Voir aussi Enrico Pizetti e Tomasso Salvemini, (dir.) *Memorie di metodologica statistica*, Roma: Libreria Eredi Virilio Veschi.
- Ginsburgh, V., et Noury, A. (2008), "The Eurovision Song Contest: Is voting political or cultural?", *European Journal of Political Economy* 24, 41-52.
- Ginsburgh, V., Ortuno-Ortin, I., et Weber, S. (2005), "Disenfranchisement in linguistically diverse societies. The case of the European Union", *Journal of the European Economic Association* 3, 946-964
- Ginsburgh, V., Ortuno-Ortin, I., et Weber, S. (2007), "Learning foreign languages. Theoretical and empirical implications of the Selten and Pool model", *Journal of Economic Behavior and Organization* 64, 946-964.
- Ginsburgh, V., Toubal, F., et Melitz, J. (2014), Foreign language learning, CEPR Discussion Paper 10101.
- Ginsburgh, V., et Weber, S. (2005), "Language disenfranchisement in the European Union", *Journal of Common Market Studies* 43, 273-286.
- Ginsburgh, V., et Weber, S. (2011), *How Many Languages Do We Need. The Economics of Linguistic Diversity*, Princeton, NJ : Princeton University Press.
- Ginsburgh, V., et Weber, S. (à paraître), "Linguistic distances and their use in economics", dans Ginsburgh, V. et Weber, S. (dir.), *The Palgrave Handbook of Economics and Language*, Houndmills Basingstoke : Palgrave Macmillan.
- -Ginsburgh, V., Weber, S., et Weyers, S. (2011) "The economics of literary translation: Some theory and evidence", *Poetics* 39, 228-246.
- Gray, R., et Atkinson, Q. (2003), "Language-tree divergence times support the Anatolian theory of Indo-European origin", *Nature* 426, 435-439.
- Greenberg, J. (1956), "The measurement of linguistic diversity", *Language* 32, 109-115.
- Hagège, C. (1996), *L'homme de paroles*, Paris : Fayard.
- Hart-Gonzalez, L., et Lindemann, S. (1993), "Expected achievement in speaking proficiency", Foreign Service Institute, Department of State: School of Language Studies.
- Heilbron, J. (1999) "Towards a sociology of translation: Book translations as a cultural world system", *European Journal of Social Theory* 2, 429-444.
- Heilbron, J., et Sapiro, G. (à paraître), "Translation : Economic and sociological perspectives", dans Ginsburgh, V., et Weber, S. (dir.), *The Palgrave Handbook of Economics and Language*, Houndmills Basingstoke : Palgrave Macmillan.
- Hombert, J-M., et Lenclud, G. (2014), *Comment le langage est venu à l'homme*, Paris : Fayard.
- Hoskins, C., McFadyen, S., et Finn, A. (1997), *Global Television and Film*, Oxford : Clarendon Press.

- Laitin, D. (2000) "What is a language community?", *American Journal of Political Science* 44, 142-155.
- Levenshtein, V. (1966), "Binary codes capable of correcting deletions, insertions, and reversals", *Cybernetics and Control Theory* 10, 707-710.
- Massey, D., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., et a Taylor, E. (1993), "Theories of international migration: A review and appraisal", *Population and Development Review* 19, 431-466.
- McMahon, A., et McMahon, R. (2005), *Language Classification by Numbers*, Oxford: Oxford University Press.
- Melitz, J. (2007), "The impact of English dominance on literature and welfare", *Journal of Economic Behavior and Organization* 64, 193-215.
- Melitz, J. (2008), "Language and foreign trade", *European Economic Review* 52, 667-699.
- Melitz, J., et Toubal, F. (2014), "Native language, spoken language, translation and trade", *Journal of International Economics* 93, 351-363.
- Renfrew, C. (1990), *Archeology and Language : The Puzzle of Indo-European Origins*, Cambridge, UK : Cambridge University Press.
- Ruhlen, M. (1997), *L'origine des langues : sur les traces de la langue mère*, Paris: Belin.
- Searls, D. (2003), "Linguistics: Trees of life and language", *Nature* 426, 391-392.
- Selten, R., et Pool, J. (1991), "The distribution of foreign language skills as a game equilibrium", dans Selten, R. (dir.), *Game Equilibrium Models*, vol. 4, Berlin: Springer.
- Swadesh, M. (1952), "Lexico-statistic dating of prehistoric ethnic contacts", *Proceedings of the American Philosophical Society* 96, 121-137.
- Swadesh, M. (1972) "What is glottochronology?" dans Swadesh, M. (dir.) *The Origin and Diversification of Languages*, London: Routledge & Kegan Paul.
- Tinbergen, J. (1962), *Shaping the World Economy : Suggestions for an International Economic Policy*, New York : The Twentieth Century Fund.

“Sur quoi la fondera-t-il l'économie du monde qu'il veut gouverner? Sera-ce sur le caprice de chaque particulier? Quelle confusion! Sera-ce sur la justice? Il l'ignore.”

Pascal

FERDi

Créée en 2003, la **Fondation pour les études et recherches sur le développement international** vise à favoriser la compréhension du développement économique international et des politiques qui l'influencent.



Contact

www.ferdi.fr

contact@ferdi.fr

+33 (0)4 73 17 75 30