Compalex : outil d'analyse dialectométrique pour une comparaison phono-lexicale synchronique des parlers d'une zone géographique

Josué Ndamba

Groupe de Recherche Langue et Informatique (GRELI) du Centre de Recherche en Sciences Sociales (CRESS), Brazzaville – Congo
et Faculté des Lettres et des Sciences Humaines, Brazzaville – Congo
Mél: jondamba@yahoo.fr

ABSTRACT

This paper presents the software "Compalex" that processes lexical data of two or several languages (or dialects) of a geographical area in view to determine the degree of intelligibility that exist between they. Softwares that exist nowadays calculate the common root percentage between languages. Thereby results show far more historical relations between the dialects or languages. Compalex processes both common root percentages between languages and the sounds that these common roots share.

Thereby, results give a more reliable indication about the way speakers of these different languages understand mutually.

This software runs under Windows 95 or later version.

1. INTRODUCTION

Dans des zones comme l'Afrique sub-saharienne où les textes des linguistes présentent une multitude de « langues », la distinction entre « langue » et « dialectes d'une langue » est un problème urgent pour réduire le problème du fractionnement dialectal et connaître le nombre réel de langues en présence dans chaque pays. Pour cela, les travaux de comparaison des différents parlers en présence s'avèrent indispensables. Mais cela implique la description linguistique préalable de chacun de ces parlers, et nécessite du coup beaucoup de temps, de personnel et de moyens financiers dont les pays africains ne disposent pas toujours. Il reste la possibilité de recourir à la comparaison lexicale pour évaluer l'intelligibilité entre les parlers en présence et procéder à des regroupements en langues de ceux qui présentent un fort degré d'intercompréhension. La méthode est rapide du fait qu'il suffit de parcourir l'espace géographique pour relever des listes lexicales et éventuellement des phrases-types puis les comparer. Elle a prouvé sa fiabilité. Mais alors le recours à l'outil informatique devient indispensable pour vaincre le « problème de masse » et réduire le taux d'erreurs dans les comparaisons à effectuer.

2. LES PROGRAMMES ACTUELS

Trois logiciels de comparaison lexicale existent à ce jour. Ils présentent chacun un certain nombre d'avantages sur les autres et beaucoup de limites.

2.1 Lexistat de Thilo Shadeberg

« *Lexistat* » de Thilo Shadeberg (Leiden) a un avantage sur les autres : il trace automatiquement le diagramme de proximité des langues ou des dialectes étudiés. La principale limite de ce programme réside dans le fait que les jugements de ressemblance à émettre pour chaque paire d'items sont d'une simplicité extrême. La seule alternative c'est de dire qu'il y a ressemblance ou non entre les deux formes comparées. Mais on ne tient pas du tout compte du degré de ressemblance entre les deux formes. A titre d'exemple, pour quatre langues L1, L2, L3 et L4 ayant les formes suivantes pour l'item "bouche":

Table 1: Exemple item « bouche » pour 4 langues

Item	Langue 1	Langue 2	Langue 3	Langue 4
bouche	munwa	monya	onu	kulu

la ressemblance entre L1 et L2 est comptabilisée de façon identique que celle entre L1 et L3 ou celle entre L2 et L3. En l'occurrence on dira qu'il s'agit de la même série de ressemblance qui sera notée par le même indice 1. L4 sera considéré comme une série distincte qui sera noté 2.

Cette conception, répétée sur une centaine de mots nous amènera à la fin de l'analyse à considérer que les pourcentages de ressemblance entre L1/L2, L1/L3 et L2/L3 sont très proche (par exemple 85, 83 et 80%).
Or, de toute évidence, il y a plus de ressemblance pour l'item "bouche" entre L1/L2 qu'entre L1/L3. Cet algorithme de comparaison développé par Carrol & Dyen [Car62] basé sur les "séries de ressemblance" présente un intérêt certain pour un travail historique, du fait que la méthode comparative cherche à identifier les séries de ressemblances partagées entre deux ou plusieurs langues et les correspondances phonétiques régulières. Mais dans une perspective dialectométrique synchronique où l'on cherche à déterminer le degré de proximité entre les parlers d'une zone, il est tout à fait insuffisant.

2.2. Word Survey de J. Wimbish

Le programme «*Word Survey* » de J. Wimbish [Wim89] présente quelques avantages sur le premier, notamment la possibilité de mesurer la divergence phonétique entre les dialectes et de faire ressortir les correspondances entre les parlers testés. Mais le modèle d'analyse pour les comparaisons reste le même que dans le programme précédent.

2.3. Le programme de W. Möhlig

Le programme de *W. Möhlig* [Möh86] constitue une première réponse à cette limite. En donnant la préférence à ce qu'il appelle le procédé qualitatif par rapport au procédé quantitatif des méthodes précédentes, il ajoute des degrés entre la divergence totale (non ressemblance entre deux formes) et l'identité totale: (a) divergence partielle morphologique, (b) divergence partielle phonologique, (c) divergence partielle accumulée (morphologique + phonologique), et il attribue des valeurs constantes à ces degrés. Ces valeurs vont de 100% pour l'identité totale à 0% pour la divergence totale; en passant par 75% (divergence partielle morphologique), 50% (divergence partielle phonologique) et 25% (divergence partielle accumulée).

S'il est vrai qu'un algorithme fondé sur ce modèle de comparaison représente un progrès certain sur celui des deux premiers programmes, il reste qu'il ne permet pas une comparaison très fiable dans le cadre d'une mesure du degré d'intelligibilité entre les parlers d'une zone, dans une optique strictement synchronique. A titre d'exemple, si les deux formes de L1 et L2 d'un item présentent deux divergences partielles phonologiques, elles seront comptabilisées de façon identique que deux formes qui présentent deux divergences morphologiques ou deux autres qui présentent deux divergences phonologiques et deux morphologiques. Par ailleurs la hiérarchisation des degrés proposée peut paraître arbitraire. Il n'est pas évident qu'une différence de préfixe, de suffixe ou d'extension dans les langues bantoues par exemple, soit moins importante pour l'intercompréhension qu'une différence phonologique. D'autant plus que souvent, le préfixe varie du fait d'un changement phonologique: un "lo" dans L1 devient "du" dans L2. Un autre inconvénient majeur de ce programme est qu'il est écrit en langage Basic, donc non compilé et très lent.

Enfin, un inconvénient majeur commun à tous ces programmes est qu'ils donnent en fin d'analyse une matrice de pourcentages de ressemblances qui sont considérés par l'analyste comme des chiffres réels. Or toute mesure statistique implique la probabilité et la prédiction. Un pourcentage lexicostatistique doit donc être considéré comme l'estimation d'une marge. Les marges représentant deux pourcentages de ressemblance peuvent se chevaucher.

Or si la valeur de ce chevauchement est assez grande, on ne peut plus dire avec certitude que les deux chiffres représentant les pourcentages, bien qu'étant différents, représentent nécessairement des degrés de relation différents. Il est donc absolument nécessaire, avant de tirer des conclusions à partir d'une matrice, de réduire tous les pourcentages qui y sont, aux différences significatives [Sim77], [Dye63]. Ce problème est bien connu des statisticiens qui ont élaboré plusieurs tests (dont le Chi²) pour mesurer la signification de la différence entre deux pourcentages.

2.4. Le point sur ces programmes

En résumé, on peut dire que tous les programmes de comparaison lexicale qui existent présentent un certain nombre de lacunes, dont les principales sont les suivantes:

- Tous sont fondés sur un algorithme de comparaison ou trop simpliste, ou pas assez affiné.
- La plupart d'entre eux ne tracent pas de façon automatique le diagramme de proximité des arbres en présence.
- Ils fonctionnent tous dans l'environnement DOS qui est de plus en plus abandonné, au profit de WINDOWS.
- Ils ne tiennent aucun compte des marges d'erreur entre les pourcentages de ressemblances donnés dans la matrice finale.

3. LE PROGRAMME COMPALEX

Le programme Compalex (pour **Compa**raison **Lex**icale) que nous nous proposons s'appuie sur un modèle de F. Manzano (Université de Rennes), qui s'inspire des travaux dialectométriques de Henri Guiter. L'idée de départ c'est d'identifier non seulement les séries de ressemblance entre les items des parlers pris deux à deux, mais en outre, de calculer pour chaque paire de mots où la ressemblance est postulée, le pourcentage de sons qui se ressemblent. On table non seulement sur la comparaison lexicale (les formes des mots) mais également sur la comparaison phonique (les sons identiques dans les mots observés). La formule de calcul est **C = (n * 100) / N,** où **C** représente le pourcentage de sons (et éventuellement de tons) que chaque mot d'un parler a en commun avec le même mot d'un autre parler; **N** le nombre total de sons (et éventuellement de tons) que le même mot présente dans les deux parlers; et **n** le nombre de sons (et éventuellement de tons) qui sont identiques dans les formes des deux parlers.

La somme de ces pourcentages pour l'ensemble d'une paire de listes, divisée par leur nombre, donne le pourcentage de ressemblance moyen entre ces deux parlers. C'est ce que Manzano appelle le *Coefficient de Proximité Linguistique* (CPL) qui lie deux parlers: **CPL = (C1+C2+C3+Cx) / x**; où **x** représente le nombre d'items de la liste de comparaison. Toutes ces moyennes

nous donnent la matrice des pourcentages de ressemblances des parlers en présence.

Cette matrice de pourcentages est à son tour traité, pour être réduite aux seules différences significatives, à partir d'une table pré-définie (que nous empruntons à G. Simons [1977]), établissant un degré de fiabilité évaluée à partir d'une estimation de la fiabilité des données linguistiques et de la longueur des listes utilisées. Une fois la matrice de pourcentages réduite à ses différences significatives, le programme calcule le *Coefficient linguistique* (CL) de chaque parler par rapport aux autres parlers auxquels il a été comparé (c'est-à-dire la somme des CPL de chaque parler divisée par le nombre de CPL). Ensuite, il calcule la *Moyenne de Proximité Linguistique* (MPL) qui rend compte de la variation linguistique totale au niveau de tous les parlers en présence. La formule de calcul est la suivante:

$$M = 1/n \sum_{i=1}^{n} x_i$$

M = MPL
x_i = CPL de la matrice
n = nombre de x_i

A partir du CL et de la MPL, le programme peut procéder à la répartition des parlers de la zone considérée en deux groupes:

- les parlers pour lesquels le coefficient linguistique (CL) est supérieur à la moyenne de Proximité Linguistique (MPL)
- les parlers pour lesquels le CL est inférieur à la MPL.

Le calcul de l'*Indice de Variation Linguistique* (IVL) permet de mesurer l'homogénéité ou la variation qui existe au sein de chaque groupe de parler. La variance minimale étant 0, plus on s'éloigne de 0 plus il y a variation au sein du groupe de parlers; et plus l'indice est faible (proche de 0), plus il y a homogénéité au sein du groupe. Cette IVL est calculée suivant la formule:

$$IVL = 1/n \sum_{i=1}^{n} (x_i - M)^2$$

x_i = CPL
M = MPL
n = nombre de CPL

Par la suite, le programme prendra chaque groupe de parlers et calculera ses CL et sa MPL et son IVL, pour faire une nouvelle répartition en sous-groupes, à partir des formules ci-dessus, jusqu'à aboutir à des communautés binaires. Il ne restera plus qu'à tracer le diagramme de proximité des parlers en présence.

3.1. Illustration

A titre d'illustration, considérons dix items de deux langues L1 et L2 et les résultats des comparaisons par les différentes méthodes ci-dessus présentées.

Par le programme de Th. Shadeberg et **de Wimbish** on obtient les résultats suivants (table 2):

Table 2: Comparaison par la méthode de Shadeberg

Item	L1	L2	Comparaison
1. Bouche	munwa	onwa	1
2. Œil	diisu	lyeso	1
3. Dent	diinu	lyeno	1
4. Bras	kooko	kwo	1
5. Nez	mbombo	liyolo	0
6. Maison	nzo	ndako	0
7. Toit	muluri	Mwanza	0
8. Racine	muza	ozye	1
9. Ecorce	kipa	epa	1
10. Eau	mamba	mayi	0
Total			**6** soit **60%**

La colonne "comparaison" de la Table 2 nous permet de faire les remarques suivantes:

(a) Les items 1 à 4 seront notés 1 point chacun parce que présentant la même racine. Il en est de même pour 8 et 9.

(b) Par contre les items 5, 6, 7 et 10 seront notés 0 parce que présentant des racines différentes.

(c) En conséquence la ressemblance entre L1 et L2 est de 6 items sur 10, soit 60% en termes de pourcentage.

Par le programme de Möhlig les résultats se présentent ainsi (Table 3):

Table 3 : Comparaison par le programme de Möhlig

Item	Langue1	Langue2	Comparaison
1 Bouche	munwa	onwa	50% (dpp)
2 Œil	diisu	lyeo	50% (dpp)
3 Dent	diinu	lyeno	50% (dpp)
4 Bras	kooko	kwo	25% (dpa)
5 Nez	mboo	liyolo	0% (dt)
6 Maison	nzo	ndako	0% (dt)
7 Toit	muludi	Mwanza	0% (dt)
8 Racine	muza	ozye	25% (dpa)
9 Ecorce	kipa	epa	50% (dpp)
10 Eau	mamba	mayi	0% (dt)

Total	**25%**

Enfin par le programme Compalex on obtient les résultats donnés plus bas (Table 4).

Table 4 : Comparaison par le programme Compalex

Item	L1	L2	Comparaison
1. Bouche	munwa	onwa	c = (3 x 100)/5 = 60
2. Œil	diisu	lyeso	c = (2 x 100)/5 = 40
3. Dent	diinu	lyeno	c = (2 x 100)/5 = 40
4. Bras	kooko	kwo	c = (3 x 100)/5 = 60
5. Nez	mbombo	liyolo	c = 0
6. Maison	nzo	ndako	c = 0
7. Toit	muluri	mwanza	c = 0
8. Racine	muza	ozye	c = 100/5 = 20
9. Ecorce	kipa	epa	c = (2 x 100)/4 = 20
10. Eau	mamba	mayi	c = 0
Total			CPL=(240)/10= **24%**

Comme on le voit, les résultats obtenus avec les deux premières méthodes sont très élevés et ne reflètent en fait que les relations historiques entre les deux parlers. Avec la méthode de Möhlig, on obtient un pourcentage beaucoup bas, plus proche de la réalité. Ce pourcentage est encore plus bas avec Compalex. Avec une liste plus longue (généralement nous travaillons sur des listes de 220 items) on obtiendrait un écart beaucoup plus important, d'autant plus que Compalex travaille sur une matrice de pourcentage réduite aux écarts significatifs, comme nous l'avons expliqué plus haut.

Voici en outre trois matrices qui illustrent la différence de calcul entre les trois programmes. La table 5 est la matrice des pourcentages des langues du groupe Teke (famille Bantu B70) produite par Wordsurv de J. Wimbish; la table 6 donne la matrice traitée par le programme de Möhlig et la table 7 celle produite par Compalex.

Table 5 : Matrice de pourcentage des parlers Teke par le programme Wordsurv (Wimbish)

94 inzinzyu

84 85 ɛngungwɛl

84 86 83 kiküküa

70 70 73 75 tɛɛ

41 40 39 42 54 ɛmbɔsi

Table 6 : Matrice de pourcentage des parlers Teke par le programme de Möhlig

91 inzinzyu

81 82 ɛngungwɛl

81 83 80 kiküküa

68 68 70 72 tɛɛ

40 38 38 40 52 ɛmbɔsi

Table 7 : Matrice de pourcentage des parlers Teke par le programme Compalex

89 inzinzyu

80 80 ɛngungwɛl

80 80 80 kiküküa

67 67 67 67 tɛɛ

38 38 38 38 51 ɛmbɔsi

4. CONCLUSION

Le programme **Compalex** que nous proposons présente sur les précédents les avantages suivants :

- Il propose une évaluation réellement synchronique de l'intelligibilité entre les parlers en présence dans une zone géographique à partir des listes lexicales ;
- Il travaille dans l'environnement Windows (toutes versions).
- Il travaille sur une matrice de pourcentage réduite aux seules valeurs significatives.
- Il trace automatiquement le diagramme de proximité des parlers de la zone considérée.

5. BIBLIOGRAPHIE

[Dye62] Dyen I. (1962), The Lexicostatistically Determined Relationships of a Language Group, International Journal of American Linguistics,n° 28, pp. 153-161.

[Möh86] Möhlig W.J.G. (1986) "Introduction à la dialectométrie synchronique", La Méthode dialectométrique appliquée aux langues africaines, Dietrich, Reimer Verlag, Berlin, pp. 15-26.

[Sim77] Simons G.F. (1977), " Tables of Significance for Lexicostatistics", Workpapers in Papua New Guinea Languages, Vol 21, pp. 75-107.

[Wim89] Wimbish J.S. (1989), Wordsurv, A Program for Analyzing Language Survey Word Lists, SIL, Dallas, Texas.