Analyses dialectométriques du lexique berbère du Rif*, Studien zur Berberologie/ Etudes Berbères,* 4, 2009: 133-150.

ANALYSES DIALECTOMETRIQUES DU LEXIQUE BERBERE DU RIF[*]

Mena Lafkioui
Università di Milano-Bicocca – Ghent University

1. Introduction aux méthodes dialectométriques

La dialectométrie est une méthode quantitative pour calculer les distances linguistiques entre des variétés linguistiques. Les méthodes dialectométriques les plus usitées peuvent être regroupées en trois catégories :

- Méthodes traditionnelles ;
- Méthodes perceptuelles ;
- Méthodes de traitement automatique.

La première méthode traditionnelle est celle de division tribale (Hoppenbrouwers & Hoppenbrouwers 2001) qui remonte au travail de Winkler (1874). Celui-ci divisait le néerlandais en frison, saxon et franconien par l'articulation de ses connaissances dialectales avec des faits historiques sur la répartition « tribale » de l'aire néerlandophone.

Les approches traditionnelles les plus connues sont celles fondées sur le concept d'isoglosse ; celle-ci étant une ligne qui sectionne une carte géographique en zones distinctes selon la variation linguistique y détectée. La classification des variétés y est déduite de l'agencement des isoglosses, des faisceaux d'isoglosses (Goossens 1969) ou des faisceaux d'isoglosses démarcatives (Stankiewicz 1957 ; Garde 1961 ; Lafkioui sous presse 2) sur la carte géolinguistique[1]. Bien que ce procédé permette une vérification des

[*] Cet article a été réalisé dans le cadre d'une recherche post-doctorale du FWO (Fonds voor Wetenschappelijk Onderzoek-Vlaanderen).

[1] Le qualificatif de *démarcatives,* adjoint au critère dialectologique classique de « faisceaux d'isoglosses » (Goossens 1969 : 54), réfère à leur valeur structurale concernant l'aspect matériel des phénomènes autant que leur distribution relative (direction et densité). Ainsi, non seulement la dimension quantitative (nombre) des isoglosses est pertinente pour la typologie de la classification, mais aussi leur dimension qualitative, c'est-à-dire leur degré d'importance. Cependant, les isoglosses non-démarcatives peuvent aussi se révéler d'une importance pour la classification, surtout lorsqu'elles permettent une évaluation postérieure

faits visualisés, il a plusieurs inconvénients dont le plus principal est le fait qu'il soit difficile de trouver des faisceaux d'isoglosses qui découpent précisément la zone géolinguistique étudiée (Kessler 1995 ; Chambers & Trudgill 1998)[2].

Le procédé de structuration géolinguistique, par contre, sert à répartir une aire géographique en fonction de la structure linguistique de ses variétés (Moulton 1960 ; Goossens 1965 ; entre autres). Les variétés disposant du même système phonémique, par exemple, font partie du même groupe géolinguistique. Cependant, les classifications adoptant cette méthode sont essentiellement de type phonologique et manquent, par conséquent, une base d'interprétation plus conjuguée avec les autres plans linguistiques[3].

Les approches perceptuelles permettent de tracer des frontières sociolinguistiques sur la base de la « conscience dialectale » des locuteurs. L'on distingue principalement deux types d'expérimentations perceptuelles, suivant l'appartenance des locuteurs à des variétés limitrophes (« Arrow method » ; Weijnen 1946, 1966 ; Rensink 1955 ; Daan & Blok 1969 ; entre autres) ou non limitrophes (Gooskens 1997, 2002 entre autres). Dans ce dernier cas, la comparaison est faite par rapport à une variété de référence, généralement la variété « standardisée ».

Les méthodes de traitement automatique sont nombreuses et considérées actuellement comme les méthodes les plus performantes pour des raisons que nous expliquerons ultérieurement. Les fondements de la dialectométrie numérique automatisée ont été établis par Séguy (1973) avec son procédé analytique pour calculer les différences linguistiques entre les variétés de la Gascogne. La comparaison y est fondée sur un algorithme qui classe les données comme identiques ou non identiques. Le total des distances mesurées entre deux variétés correspond à leur distance linguistique. La visualisation des analyses de classification y est réalisée moyennant des lignes de type divers (grasses/non grasses, pointillées/non pointillées, marquées/non marquées) qui répartissent la région selon les différences linguistiques des variétés. De manière homologue, Goeble (1982, 1993) a calculé les ressemblances entre les variétés de l'Italie et de la Suisse

---

des résultats. Sur le rapport entre « structuralisme » et « dialectologie », voir entre autres Forquet (1956), Weinreich (1954), Grosse (1960) et Martinet (1972).

[2] L'on reproche aussi à cette méthode de ne pas pouvoir exclure complètement une certaine subjectivité, car les isoglosses seraient généralement choisies, a priori, selon les frontières linguistiques qu'elles permettent de dégager (Goossens 1977).

[3] Même si la fréquence des variantes comparées est prise en considération (Kocks 1970), cette approche ne semble pas être la plus adéquate (Heeringa 2004 : 24-25).

Méridionale. Bien que les résultats issus du calcul de Séguy et de Goeble aient le mérite d'être objectifs, ils manquent un certain raffinement dû à leur technique qui exclut toute graduation de distance.

Les procédés automatisés, s'appuyant sur la fréquence de la variante linguistique, sont fondamentalement celui de « Corpus Frequency Method » (Hoppenbrouwers & Hoppenbrouwers 1988, 2001) et de « Frequency per Word Method » (Nerbonne & Heeringa 1998, 2001). Le principe de base du premier procédé est que le degré de différence/ressemblance entre deux variétés découle de la comparaison de la fréquence des traits linguistiques marqués de leurs variantes. Le problème qui se pose pour cette approche, c'est que l'entité « mot » n'y est pas considérée comme une unité linguistique. Cet obstacle est, par contre, enlevé par la seconde approche qui accorde aux mots le statut d'unités fonctionnant en tant que tel. Cependant, les deux outils de classification ne tiennent pas compte de l'ordre des unités phoniques dans la séquence.

Ce qui distingue la mesure de distance « Levenshtein » (Lv) – et la rend plus adéquate par rapport aux autres méthodes numériques – est le fait qu'elle offre la possibilité d'intégrer dans la classification le paramètre d'agencement séquentiel des unités phoniques. Cet outil a été introduit dans la dialectométrie par Kessler (1995) qui l'a appliqué sur un corpus du gaélique d'Irlande. La mesure Levenshtein correspond à la valeur numérique du coût le plus bas des opérations (insertions, suppressions et substitutions) nécessaires pour transformer une chaîne de caractères dans une autre (Kruskal 1999). Parmi les techniques de comparaison les plus employées, figure la technique de « phone string comparison » dans laquelle toutes les opérations ont le même coût, quel que soit le degré d'affinité entre les unités phoniques : la paire [t, d] a le même coût que les paires [u, t] et [u, uː], par exemple. Dans la technique de « feature string comparison », en revanche, ce sont les traits phonétiques des unités phoniques qui sont comparés : le coût des paires [u, t] et [u, uː] n'est pas le même en raison de l'affinité phonétique plus grande entre les unités phoniques de [u, uː] par rapport à celles de [u, t].

2. Analyses dialectométriques du lexique berbère du Rif

Des différentes méthodes utilisées, nous préconisons celles assistées par l'ordinateur, car elles permettent de manier des corpus de données étendus avec une certaine commodité, tout en garantissant la précision et la cohérence des analyses, par le fait que :

- Les distances et les fréquences sont mesurées de manière automatique.

- Les données sont classifiées numériquement.
- La cartographie peut être assistée par l'ordinateur.
- Des analyses statistiques et non statistiques peuvent être effectuées et visualisées automatiquement.

Les analyses dialectométriques que nous présenterons dans cet article ont été effectuées avec le logiciel informatique libre de Kleiweg (RuG/L04)[4].

Afin d'accomplir une analyse dialectométrique visualisée, toutes les étapes de la procédure résumée ci-dessous sont indispensables (Lafkioui, sous presse 1):

Tableau 1 : Procédure générale de l'analyse dialectométrique automatisée

Etape 1	Atlas linguistique = source des données géoréférenciées
Etape 2	Matrice des données
Etape 3	Matrice des distances
Etape 4	Analyses
Etape 5	Visualisation

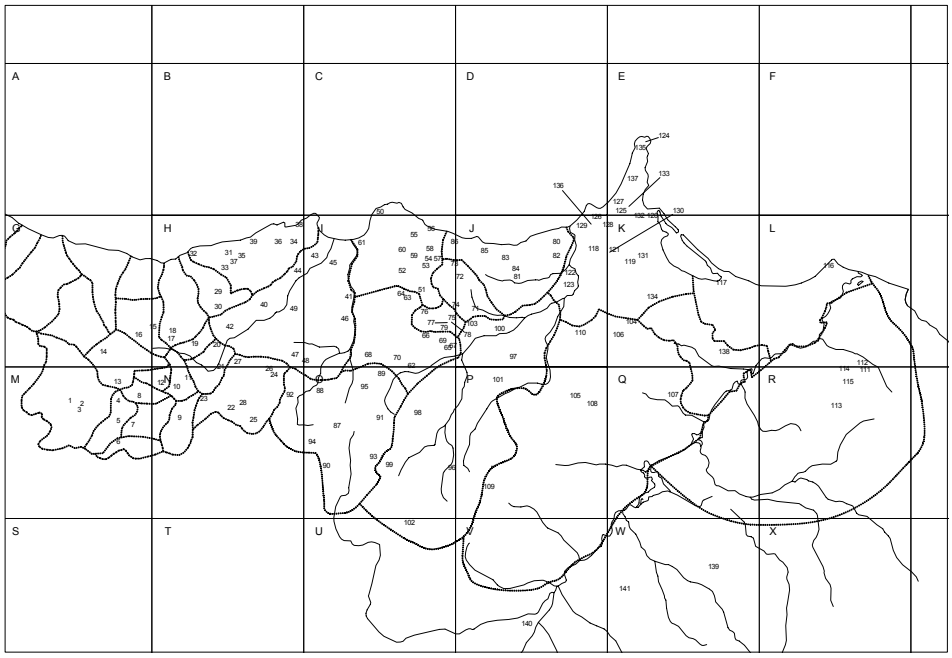
2.1. L'Atlas linguistique du Rif comme source des données

Les matériaux lexicaux comparés et classifiés dans cette étude proviennent de l'*Atlas linguistique des variétés berbères du Rif* (Lafkioui, 2007), dorénavant ALR. Il s'agit d'un corpus numérique de soixante-deux lexèmes portant sur le corps humain (cartes 295 à 315), les liens de parenté (cartes 316 à 321), les animaux (cartes 322 à 327), les couleurs (cartes 328 et 329), les numéraux (cartes 330 à 332) et un sous-ensemble de noms et verbes divers (cartes 333 à 356). Parmi ces lexèmes, onze ne disposent que d'une seule variante par variété ; les cinquante et un autres lexèmes exposent tous une co-occurrence de multiples variantes pour chaque lexème.

En raison de la réalisation automatisée de l'ALR, les données qui en sont extraites sont déjà en format numérique, ce qui a évité un gros travail de numérisation. Cependant, une conversion adaptative au logiciel RuG/L04 (Kleiweg) a été nécessaire. L'ALR offre, en outre, une carte géographique numérique précise de la région du Rif (voir carte des points d'enquête géoréférenciées du Rif, Figure 1). Cette carte est essentielle à la visualisation des analyses, exception faite du dendrogramme.

[4] Http://odur.let.rug.nl/~kleiweg/L04.

Figure 1 : Carte des points d'enquête géoréférenciés du Rif (Lafkioui 2007)



Cent quarante et un points géoréférenciés – appartenant à trente-deux tribus rifaines – ont été sélectionnés d'un ensemble de quatre cent cinquante-deux localités du Rif selon leur degré de variation linguistique (Lafkioui 2007)[5].

2.2. Matrice des données lexicales berbères du Rif

La matrice des données est composée des matériaux lexicaux numériques extraits de l'ALR (Lafkioui 2007) et convertis suivant le format du logiciel RuG/L04 (Kleiweg). En voici un petit échantillon en format numérique de l'ALR (format de Mapinfo Professional ; Tableau 2) et en format texte du logiciel RuG/L04 (Tableau 3) :

[5] Les points d'enquête ont été sélectionnés suivant le principe d'équidistance divisant le terrain d'enquête en plusieurs mailles dont chacune d'elle a été affectée d'un point qui pouvait correspondre sur le terrain à une localité. Plus la variation était grande plus les mailles ont été réduites. Les quatre cent cinquante-deux localités retenues pour cette recherche ont été, en majeure partie, choisies de façon qu'elles puissent, a priori, indiquer des frontières linguistiques, ce qui découlait principalement de la connaissance empirique et scientifique de l'enquêteur des différentes variétés parlées au Rif.

Tableau 2 : Extrait des données en format numérique de l'ALR

SECTOI	TRIBE	FULL_NAME_ND	LF461	LF462	LF463	LF464	LF465	LF466	LF467	LF468	LF469	LF470
1	Ktama	Asammer	31	51	52	32	54	51	32	33	53	52
1	Ktama	Lmexzen	31	51	52	32	54	51	32	33	53	52
1	Ktama	Ssahel	31	51	52	32	54	51	32	33	53	52
2	Taghzut	Lqel'a	31	54	52	32	54	51	34	33	54	52
2	Taghzut	Ssaqya	31	54	52	32	54	51	34	33	54	52
3	Ayt Bucibet	Tarya	31	54	52	32	54	51	32	33	54	52
4	Ayt Hmed	Mazuz	31	51	52	32	54	51	32	33	54	52
5	Ayt Bunsar	Luta	31	32	52	32	54	51	32	33	51	52
6	Ayt Bcir	Tizirt	31	51	52	32	52	51	33	31	51	52
7	Zerqet	Aghennuy	31	32	52	32	52	51	33	31	51	52
7	Zerqet	Wersan	31	32	52	32	52	51	33	31	51	52
8	Ayt Xennus	A'raben	31	32	52	32	54	51	32	33	53	52
9	Ayt Seddat	Azila	31	32	52	32	54	51	32	33	53	52
9	Ayt Seddat	Tamadda	31	32	52	32	54	51	32	33	53	52
A	Ayt Gmil	Azru n tili	53	53	53	32	51	54	31	52	53	53
A	Ayt Gmil	Tizi	53	53	53	32	51	54	31	52	53	53
B	Ayt Bufrah	Igzennayen	53	32	52	32	51	13	31	52	53	52
B	Ayt Bufrah	Iharunen	53	32	52	32	51	13	31	52	53	52
C	Targist	Ayt 'Azza	54	52	53	53	51	12	53	52	53	53
D	Ayt Mezduy	Bni Budjay	53	53	53	32	51	54	31	52	53	53
D	Ayt Mezduy	Bu'di	53	32	52	32	51	51	31	52	53	52

Tableau 3 : Extrait des données en format texte du logiciel RuG/L04

: Asammer
- aqnin
: Lmexzen
- aqnin
: Ssahel
- aqnin
: Lqel'a
- aqnin
: Ssaqya
- aqnin
: Tarya
- aqnin
: Mazuz
- aqnin
: Luta
- aqnin
: Tizirt
- aqnin
: Aghennuy
- aqnin
: Wersan
- aqnin
: A'raben
- aqnin
: Azila
- aqnin
: Tamadda
- aqnin
: Azru n tili
- aqenni
: Tizi
- aqenni
: Igzennayen
- aqenni
: Iharunen
- aqenni
: Ayt 'Azza
- aqnenniy
: Bni Budjay
- aqenni
: Bu'di
- aqenni
: Aghir Hmed
- aqenni
: Asammar
- aqenni
: Ayt Hmed
- aqenni
: Sidi Bucetta
- aqenni
: Tazrut
- aqenni
: Ufis
- aqenni
: Wad Mahkim
- aqenni
: L'ars
- aqenni
: Tufist-Imuruten
- aqenni

2.3. Matrice des distances pour le lexique berbère du Rif

Dans cette section, nous mettrons en contraste les trois techniques de comparaison numérique les plus usitées : la mesure de distance binaire (algorithme de Hamming), la mesure de distance « Gewichteter Identitätswert » (identité pondérée), et la mesure de distance Levenshtein. Nous les appliquerons sur le lexique berbère du Rif afin de tester leur validité et d'en sélectionner la plus appropriée au berbère. Chaque mesure de distance permet d'obtenir des valeurs numériques précises issues de la comparaison linguistique entre les variétés du Rif. Ces valeurs composent les matrices de distance (matrices symétriques N x N, N = somme des variétés) dont la configuration diverge selon l'algorithme adopté.

2.3.1. La mesure de distance binaire

La mesure binaire (Bin) permet de classifier les unités lexicales comme étant identiques ou non identiques. Le rapport de comparaison est donc de 0-1 ; 0 = ressemblance et 1 = différence. Voici un extrait de la matrice de distance binaire pour le lexème « talon » (ALR, carte 312) :

Tableau 4 : Extrait de la matrice de distance binaire pour le lexème « talon »

	Tizirt	Aghennuy	Wersan	A'raben	Azila	Tamadda	Azru n tili	Tizi
Wersan	0	0	0	0	0	0	1	1
A'raben	0	0	0	0	0	0	1	1
Azila	0	0	0	0	0	0	1	1
Tamadda	0	0	0	0	0	0	1	1
Azru n tili	1	1	1	1	1	1	0	0
Tizi	1	1	1	1	1	1	0	0
Igzennayen	1	1	1	1	1	1	0	0
Iharunen	1	1	1	1	1	1	0	0
Ayt 'Azza	1	1	1	1	1	1	1	1

2.3.2. La mesure de distance « Gewichteter Identitätswert »

La mesure « Gewichteter Identitätswert » (GIW) diverge de la mesure binaire par le fait que la fréquence de la variante lexicale joue également un rôle dans la comparaison : les variantes à fréquence basse pèsent plus lourd que les variantes à fréquence élevée. Les valeurs de distance obtenues par cette technique sont comprises entre 0 et 1, soit $\{0 \leq d \leq 1\}$. En voici un exemple provenant de la matrice de distance du lexème « talon » :

Tableau 5 : Extrait de la matrice de distance GIW pour le lexème « talon »

	Tizirt	Aghennuy	Wersan	A'raben	Azila	Tamadda	Azru n tili	Tizi
Wersan	0.0501792	0.0501792	0	0.0501792	0.0501792	0.0501792	1	1
A'raben	0.0501792	0.0501792	0.0501792	0	0.0501792	0.0501792	1	1
Azila	0.0501792	0.0501792	0.0501792	0.0501792	0	0.0501792	1	1
Tamadda	0.0501792	0.0501792	0.0501792	0.0501792	0.0501792	0	1	1
Azru n tili	1	1	1	1	1	1	0	0.215054
Tizi	1	1	1	1	1	1	0.215054	0
Igzennayen	1	1	1	1	1	1	0.215054	0.215054
Iharunen	1	1	1	1	1	1	0.215054	0.215054
Ayt 'Azza	1	1	1	1	1	1	1	1

2.3.3. La mesure de distance Levenshtein

Les valeurs de distance qui dérivent de la comparaison fondée sur la mesure Levenshtein – algorithme qui tient compte de l'ordre séquentiel des unités phoniques dont les lexèmes sont composés – varient entre 0 et 1, ($\{0 \leq d \leq 1\}$), comme le montre l'extrait suivant :

Tableau 6 : Extrait de la matrice de distance Lv pour le lexème « talon »

	Tizirt	Aghennuy	Wersan	A'raben	Azila	Tamadda	Azru n tili	Tizi
Wersan	0	0	0	0	0	0	0.6	0.6
A'raben	0	0	0	0	0	0	0.6	0.6
Azila	0	0	0	0	0	0	0.6	0.6
Tamadda	0	0	0	0	0	0	0.6	0.6
Azru n tili	0.6	0.6	0.6	0.6	0.6	0.6	0	0
Tizi	0.6	0.6	0.6	0.6	0.6	0.6	0	0
Igzennayen	0.6	0.6	0.6	0.6	0.6	0.6	0	0
Iharunen	0.6	0.6	0.6	0.6	0.6	0.6	0	0
Ayt 'Azza	0.555556	0.555556	0.555556	0.555556	0.555556	0.555556	0.111111	0.111111

Ces valeurs sont le résultat de la sélection du calcul le moins coûteux pour transformer une unité lexicale – en tant que chaîne d'unités phoniques – en une autre. Le tableau suivant expose les coûts les plus bas des opérations permettant de modifier les chaînes de caractères de *awrez* (talon) en *inerz* (talon) :

Tableau 7 : Coûts des opérations permettant de modifier *awrez* en *inerz* (talon)

		a	**w**	**r**	**e**	**z**
	0	0.5	1	1.5	2	2.5
i	0.5	1	1.5	2	2.5	3
n	1	1.5	2	2.5	3	3.5
e	1.5	2	2.5	3	2.5	3
r	2	2.5	3	2.5	3	3.5
z	2.5	3	3.5	3	3.5	3

Le coût le moins élevé des opérations modifiant *awrez* en *inerz* est 3, ce qui fait que la distance entre ces deux lexèmes est de 3/5 (5 étant le total des caractères) ; soit, la distance de Levenshtein est de 60 %. Ces calculs sont basés sur des opérations qui coûtent 0.5 pour une insertion ou une suppression et 1 pour une substitution. Ainsi, par exemple :

Tableau 8 : Exemple du calcul de distance Lv pour modifier *awrez* en *inerz* (talon)

Tamadda	a	w	r	e		z	
Tizi	i	n		e	r	z	
Distance Lv	1	1	0,5	0	0,5	0	**3/5 * 100 = 60 %**

2.4. Analyses dialectométriques numériques du lexique berbère du Rif

A partir des matrices de distance, des analyses de comparaison numériques du lexique berbère peuvent être réalisées par le biais de deux types de technique : « Cluster analysis » (analyse de regroupement) et « Multidimensional scaling » (analyse de graduation multidimensionnelle). La technique de « Cluster analysis » (CA) consiste à regrouper les données par réduction de la matrice de distance moyennent des algorithmes variés. A la suite de Kleiweg (RuG/L04), nous avons adopté l'algorithme de Ward (variance minimum) qui est généralement considéré comme un des algorithmes les plus adéquats. Multidimensional scaling (MDS), par contre, est :

> […] a technique that, using a table of differences, tries to position a set of elements into some space, such that the relative distances in that space between all elements corresponds as close as possible to those in the table of differences. (Kleiweg, RuG/L04)[6].

2.5. Visualisation des analyses dialectométriques du lexique berbère du Rif

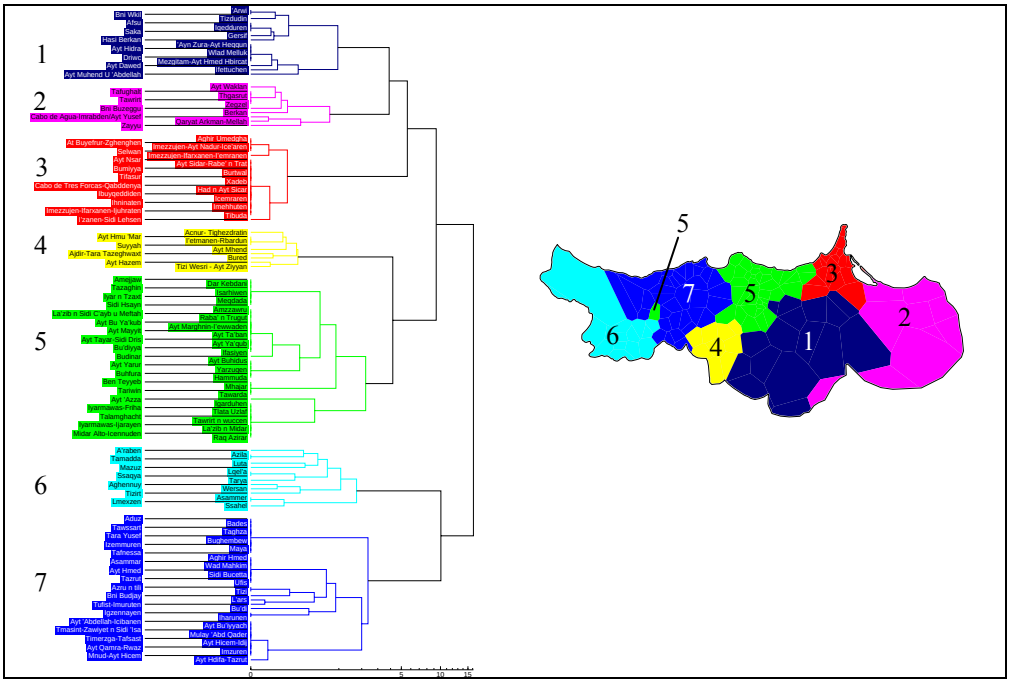
La classification par regroupement (CA) a, pour être visualisée, nécessairement recours à un dendrogramme, une sorte d'arborescence complexe, généralement en couleur, dont les branches représentent les variétés. Ce dendrogramme peut être associé avec une carte géographique numérique, ce qui donne comme résultat une carte géolinguistique qui montre la répartition des variétés selon les différences linguistiques et les critères de classification retenus. Les analyses par graduation multidimensionnelle (MDS), en revanche, offre directement des cartes où la variation relative est représentée de façon graduelle par des nuances de couleurs différentes.

[6] […] une technique qui, au moyen d'un tableau de différences, tente de positionner un ensemble d'éléments dans l'espace, de telle manière que les distances relatives entre les éléments de cet espace correspondent autant que possible à celles du tableau des différences.

2.5.1. Visualisation et interprétation des analyses CA

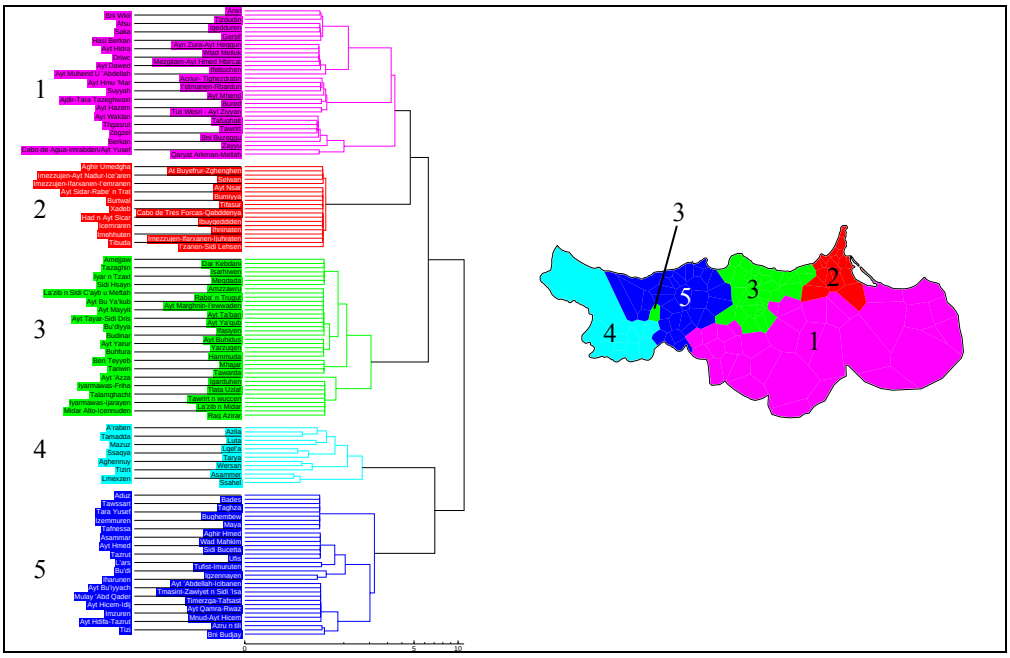
La structure hiérarchique du dendrogramme et la répartition des variétés sur la carte du Rif qui y est associé varient sensiblement selon l'algorithme de distance (Bin, GIW ou Lv) appliqué. Ainsi, l'on aperçoit pour la mesure binaire (Figure 2) une structure composée de sept groupes principaux, regroupés dans deux sous-groupes : le sous-groupe mineur contenant les groupes 6 et 7, et le sous-groupe majeur contenant les groupes 1 à 5 ; la distance entre ces deux sous-groupes étant de 16.17. Cette valeur de distance relativement élevée indique une frontière linguistique nette après le groupe 7 qui est délimité à droite par les variétés des Ayt Weryaghel et des Ayt 'Ammart. Le sous-groupe majeur connaît une subdivision assez équilibrée (d = 9.34) entre d'une part les groupes 4 et 5 (variété de Targuist incluse) et d'autre part les groupes 1 à 3 qui, à leur tour, ont également subi une subdivision. La seconde frontière linguistique importante coïncide donc avec les variétés limitrophes des groupes 4 (Igzennayen) et 5 (Ayt S'id et Ayt Tuzin).

Figure 2 : Dendrogramme vs Carte CA - Bin – Tout le lexique



La classification fondée sur l'algorithme GIW diverge considérablement de celle fondée sur l'algorithme Bin, car elle aboutit à un ensemble de cinq regroupements (Figure 3), dont le regroupement 1 englobe les sous-groupes 1, 2 et 4 de la classification Bin (Figure 2). Cependant, la frontière linguistique principale détectée par le biais de GIW – frontière tracée après les variétés du groupe 5 – est identique à celle que dégage le dendrogramme Bin, bien que la distance entre les deux sous-ensembles principaux soit moins élevée pour GIW (d_{GIW} = 10.87) que pour Bin (d_{Bin} = 16.17), en raison de l'intégration de la fréquence dans la comparaison.

Figure 3 : Dendrogramme vs Carte CA - GIW – Tout le lexique



La classification issue des analyses dialectométriques fondées sur la mesure de distance Lv résulte dans une configuration asymétrique de 7 groupes répartis en 2 sous-ensembles distanciés l'un de l'autre de 8.08 (Figure 4). Ce dendrogramme partage la même délimitation linguistique prépondérante (entre groupes 6 et 3-4) avec les autres dendrogrammes. Ce constat est corroboré par les cartes CA_{Lv} présentées dans la Figure 5, dont celle à deux regroupements indique clairement la frontière linguistique la plus distinctive. Il importe de remarquer que la carte CA_{Lv} (Figure 4) affiche une répartition des variétés analogue à celle de la carte CA_{Bin}, bien que la composition de leur dendrogramme respectif soit divergente.

Figure 4 : Dendrogramme vs Carte CA – Lv – Tout le lexique

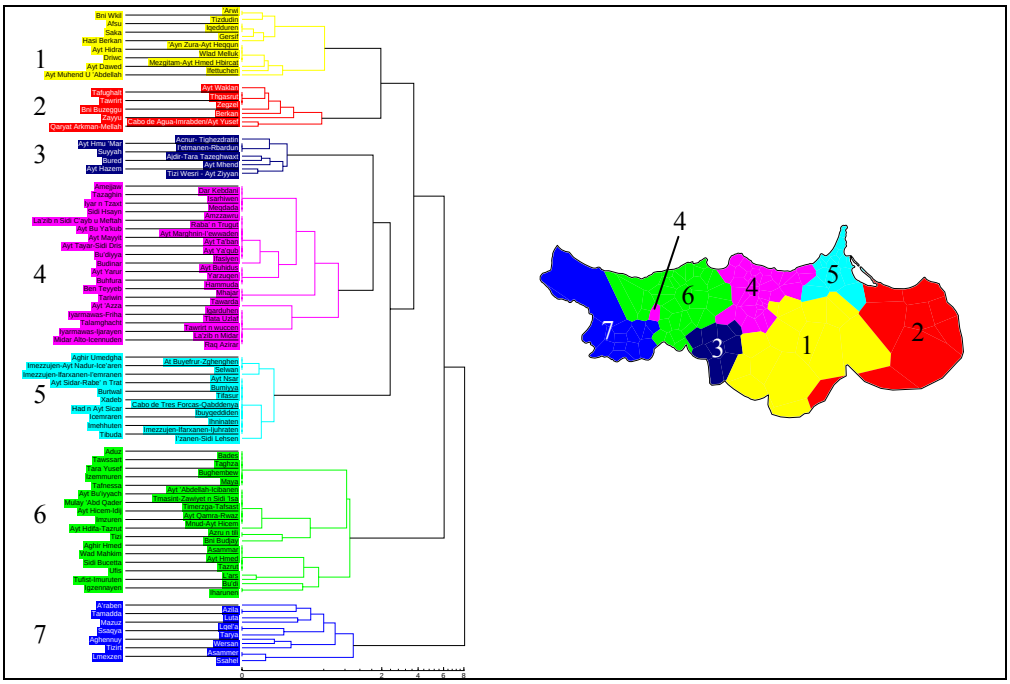
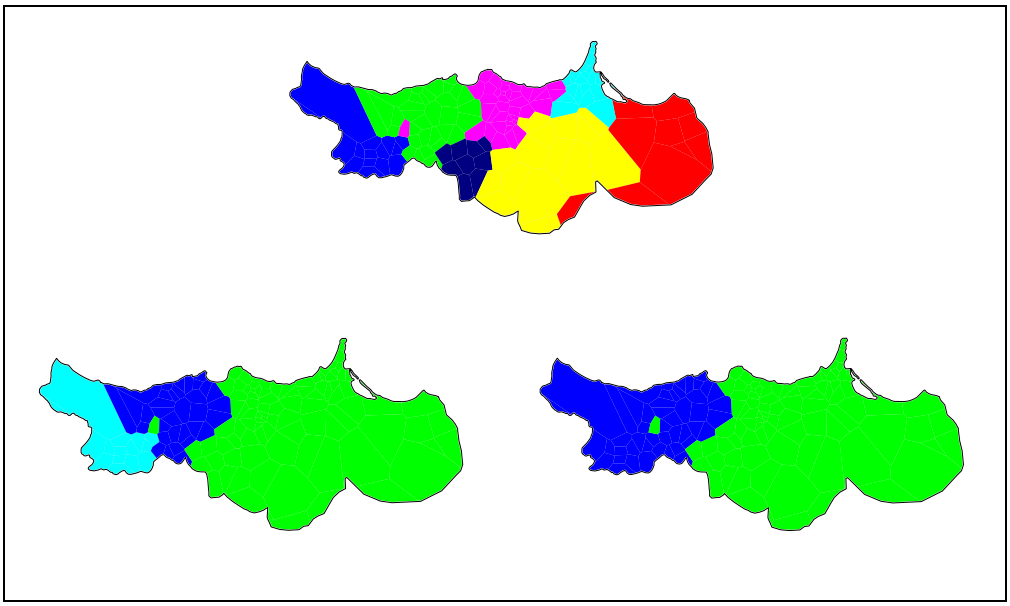


Figure 5 : Cartes CA_{Lv} – 7 groupes vs 3 groupes vs 2 groupes - Tout le lexique

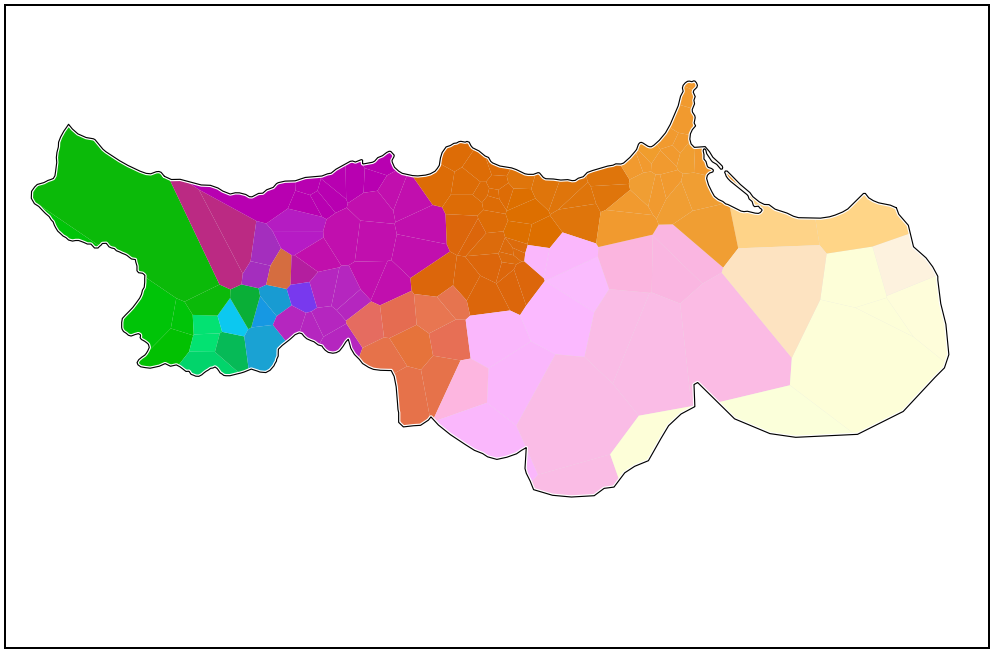


2.5.2. Visualisation et interprétation des analyses MDS

La technique MDS a l'avantage majeur de garantir une objectivité et une précision lors de la phase d'analyse des matériaux, parce qu'elle en exclut tout paramétrage externe. Ainsi, par exemple, l'on n'y peut pas modifier le nombre de regroupements. C'est le système d'analyse qui le fournit automatiquement. Chaque variété y a sa propre couleur. Ce sont les contrastes de couleurs qui servent à l'interprétation des données linguistiques comparées : une continuité de couleur indique une corrélation parfaite entre les lexèmes, alors qu'une mosaïque de couleurs dévoile une corrélation faible entre eux.

L'aire du Rif connaît, par le biais de MDS, une répartition en 7 grandes zones, quelle que soit la mesure de distance appliquée (Figure 6). La répartition des variétés sur les cartes MDS est quasi similaire pour Bin et GIW ; seulement quelques différences mineures de nuances de certaines couleurs ont été observées. La carte MDS_{Lv} ressemble fortement aux deux autres ; la seule distinction significative constatée est l'apparition d'une petite subdivision a l'intérieur du groupe des variétés occidentales.

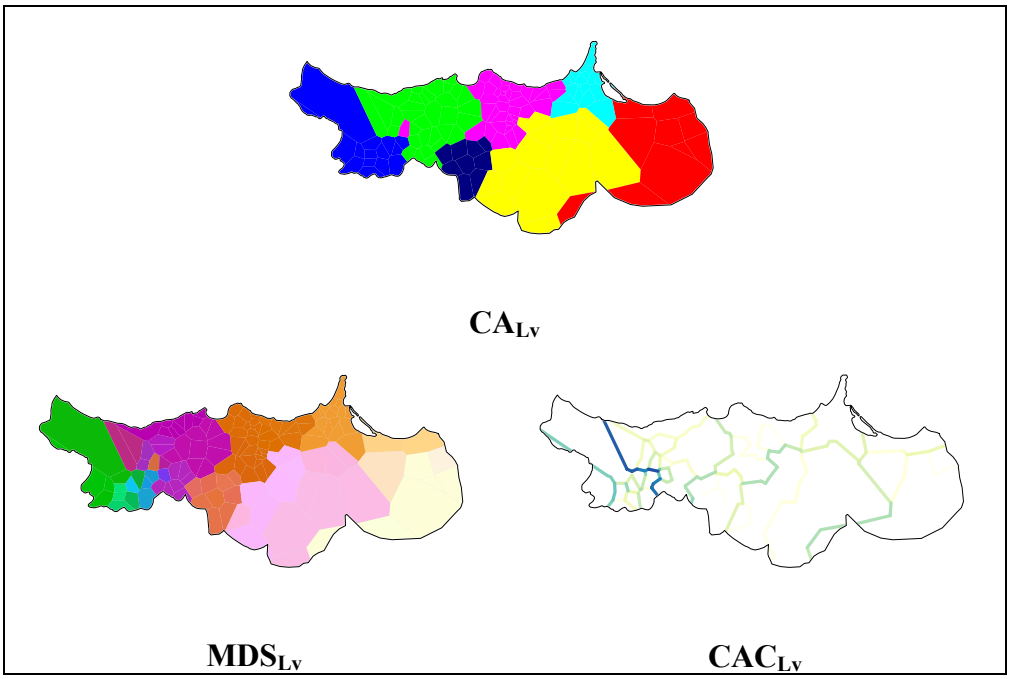
Figure 6 : Carte MDS_{Lv} – Tout le lexique



3. Résultats comparatifs

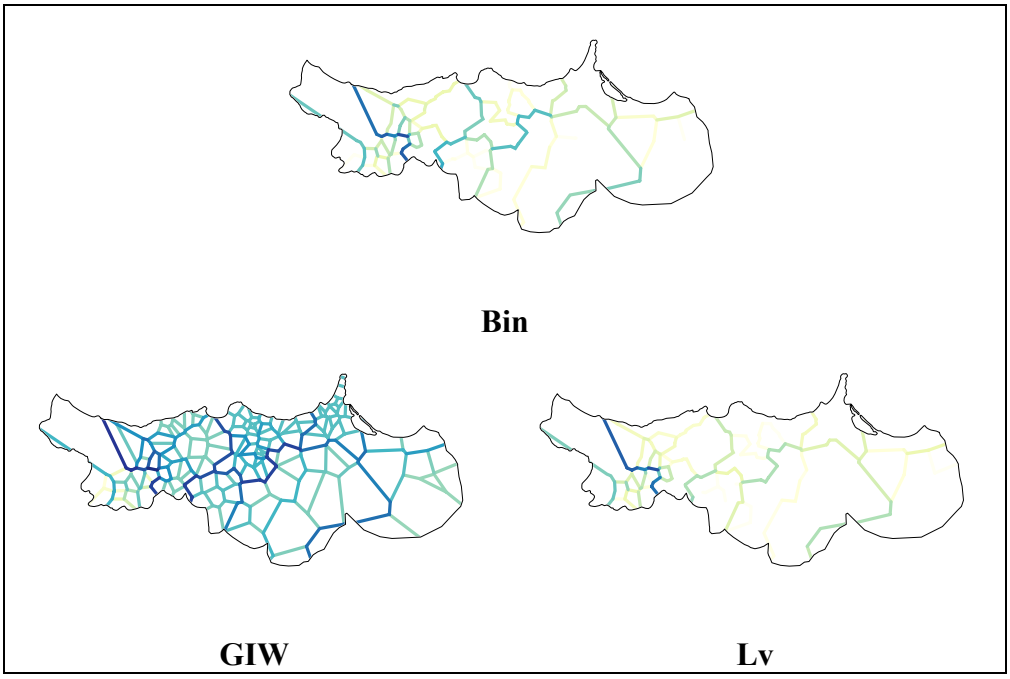
Du fait de son exactitude, le procédé MDS est le plus approprié pour l'analyse dialectométrique du lexique berbère. Il forme, de ce fait, un critère par rapport auquel les autres méthodes dialectométriques peuvent être mises en contraste. Parmi les classifications par regroupement (CA), les classifications CA_{Bin} et CA_{Lv} rejoignent le mieux la distribution affichée par les cartes MDS (7 groupes). La classification CA_{Lv} montre, en outre, un raffinement davantage élevé du fait qu'elle prend en considération la variation phonique des occurrences autant que leur agencement dans les lexèmes. Cependant, toute analyse fondée sur la mesure de distance Lv (CA aussi bien que MDS) ignore la hiérarchie qui existe entre les unités phoniques (unités phonétiques = unités phonémiques), à moins qu'on leur accorde des poids divers moyennant un paramétrage spécifique, ce qui revient à construire un système phonologique à l'intérieur du logiciel, impliquant un travail laborieux et trop coûteux par rapport aux profits.

Figure 7 : Cartes CA_{Lv} vs MDS_{Lv} vs CAC_{Lv}



La classification par regroupement (CA) à l'avantage d'indiquer précisément les frontières linguistiques significatives. Les cartes CAC (Composite cluster map = carte de regroupement composite ; Figures 7 et 8) les marquent par des lignes à couleur foncée. Comparé aux délimitations distinctives dégagées par les dendrogrammes et les cartes CA correspondantes des Figures 2 à 5, la frontière linguistique principale de la carte CAC de la Figure 7 est tracée encore davantage vers l'ouest. Il importe, cependant, de signaler que les cartes CAC ne nous paraissent pas les plus aptes à visualiser la classification du lexique berbère du Rif, en raison de la difficulté d'interprétation des données, dû à leur représentation assez chaotique (Figure 8)[7].

Figure 8 : Cartes CAC – Bin vs GIW vs Lv – Tout le lexique



[7] Kleiweg (RuG/L04) propose certaines alternatives pour l'algorithme Ward qui semble être à la base de ce désordre visuel par lequel les cartes CA_{GIW} sont les plus touchées.

Références bibliographiques

Chambers J. K. & Trudgill, P. 1998. *Dialectology*. Cambridge University Press, Cambridge, 2nd edition.

Daan, J. & Blok, D. P. 1969. *Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde*, volume XXXVII of Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam. Noord-Hollandsche Uitgevers Maatschappij, Amsterdam.

Forquet, J. 1956. Linguistique structurale et dialectologie, *Festgabe Frings*, 190-203.

Garde, P. 1961. Réflexions sur les différences phonétiques entre les langues slaves, *Word*, XVII : 34-62.

Goebl, H. 1982. *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*, Philosophisch-Historische Klasse Denkschriften, volume 157, Verlag der Osterreichischen Akademie der Wissenschaften, Vienna.

Goebl, H. (1993). Probleme und Methoden der Dialektometrie: Geolinguistik in globaler Perspektive. In: Viereck, W. (ed.), *Proceedings of the International Congress of Dialectologists*, , Stuttgart. Franz Steiner Verlag, volume 1, 37–81

Goossens, J. 1965. *Die niederländische Strukturgeographie und die "Reeks Nederlandse Dialectatlassen*", Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam, volume XXIX , N.V. Noord-Hollandsche Uitgevers Maatschappij, Amsterdam.

Goossens, J. 1969. *Strukturelle Sprachgeographie. Eine Einführung in Methodik und Ergebnisse*. – Heidelberg.

Grosse, R. 1960. Strukturalismus und Dialektgeographie. – *Biuletyn Fonograficzny*, III : 89-101.

Heeringa, W. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD. Dissertation, Rijksuniverstiteit Groningen, Groningen.

Hoppenbrouwers, C. & Hoppenbrouwers, G. 1988. De feature frequentie-methode en de classificatie van Nederlandse dialecten. – *TABU, Bulletin voor taalwetenschap*, 18(2) : 51-92.

Hoppenbrouwers, C. & Hoppenbrouwers, G., 2001. *De indeling van de Nederlandse streektalen : Dialecten van 156 steden en dorpen geklasseerd volgens de FFM (feature frequentie methode).* – Koninklijke Van Gorcum, Assen.

Kessler, B. 1995. Computational dialectology in Irish Gaelic. In: *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, Dublin. EACL, 60–67.

Kruskal, J. 1999. An overview of sequence comparison. – In : D. Sankoff & J. Kruskal (eds.), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison-Wesley, MA, 1-44.

Lafkioui, M. 2007. *Atlas linguistique des variétés berbères du Rif,* Köln, Rüdiger Köppe Verlag (in Berber Studies, volume 16), 2007, 291 p. (format « A3 », in color, 356 maps + 47 tables).

Lafkioui, M. sous presse 1. Etudes de géographie linguistique berbère : variation géolinguistique et classification dialectométrique, *Le Bulletin des Séances de l'Académie des Sciences d'Outre-Mer*, 22 p.

Lafkioui, M. sous presse 2. Pour la recherche dialinguistique du berbère. Le cas du tarifit. In: A. El Aissati (ed.), *From Oral Discourse Analysis in Berber to Academic Language Skills in Dutch*, Harrossowitz, 9 p.

Martinet, A. 1972. Structural dialectology, *Pakha Sanjam*, N° Spécial.

Moulton, W. G. 1960. The short vowel systems of northern Switzerland. *Word*, 16 : 155–182.

Nerbonne, J. & Heeringa, W. 1998. Computationale vergelijking en classificatie van dialecten. – *Taal en Tongval*, 50(2) : 164-193.

Nerbonne, J. & Heeringa, W. 2001. Computational Comparison and Classification of Dialects. – *Dialectologia et Geolinguistica*, 9 : 69-83.

Rensink, W. G. 1955. Dialectindeling naar opgaven van medewerkers. *Mededelingen der Centrale Commissie voor Onderzoek van het NederlandseVolkseigen*, 7: 20–23.

Séguy, J. 1973. La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique romane*, 37: 1-24.

Stankiewicz, E. 1957. On discreteness and Continuity in Structural dialectology, *Word*, 13: 44-59.

Weijnen, A. (1946). De grenzen tussen de oost-noord-Brabantse dialecten onderling. In: *Oost-Noordbrabantsche dialectproblemen: lezingen gehouden voor de dialectencommissie der Koninklijke Nederlandsche Akademie van Wetenschappen op 12 april 1944,* Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam, volume VIII, 1-15. Noord-Hollandsche Uitgevers Maatschappij, Amsterdam.

Weijnen, A. (1966). *Nederlandse dialectkunde*. Studia Theodisca. Van Gorcum, Assen.
Weinreich, U. 1954. Is a structural dialectology possible ?, *Word*, 10 : 388-400.
Winkler, J. 1874. *Algemeen Nederduitsch en Friesch Dialecticon*. Martinus Nijhoff, 's-Gravenhage.