Un usage particulier de l'algorithme de Damerau-Levenshtein dans le domaine occitan

Guylaine Brun-Trigaud

▶ **To cite this version:**

Guylaine Brun-Trigaud. Un usage particulier de l'algorithme de Damerau-Levenshtein dans le domaine occitan. 2013. halshs-01067295

HAL Id: halshs-01067295
https://shs.hal.science/halshs-01067295v1

Preprint submitted on 23 Sep 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Un usage particulier de l'algorithme de Damerau-Levenshtein dans le domaine occitan

Guylaine Brun-Trigaud
Univ. Nice Sophia Antipolis,
CNRS BCL, UMR 7320,
06300 Nice, France

1. Introduction

Après avoir terminé la saisie des dernières cartes des atlas linguistiques régionaux, édités par le CNRS, qui constituent la source essentielle du Thesaurus Occitan (THESOC) (près de 800000 data disponibles actuellement en ligne[1]), les responsables scientifiques du projet, Jean-Philippe Dalbera et Michèle Oliviéri, m'ont confié la mission de compléter ce corpus avec les données inédites de l'ALLOc et de l'ALLOr[2].

Jusqu'ici la saisie dans le THESOC s'était toujours accomplie à partir des cartes publiées, avec la possibilité de faire usage (plus ou moins) du copier-coller, notamment dans un domaine comme celui de l'ALLOc qui est relativement homogène[3]. Or cette solution ne paraissait plus possible puisque la saisie devait s'effectuer à partir des carnets d'enquêtes (il est impossible d'avoir en même temps sous les yeux, même sous la forme d'images numérisées, les 131 carnets pour consulter les réponses de chaque point pour une même notion ...).

Afin de gagner du temps et ne pas avoir à entreprendre la longue tâche qui aurait consisté à reprendre une à une, pour les 131 points d'enquête, les 430

[1] http://thesaurus.unice.fr/.

[2] Le CNRS a interrompu en 1996 la publication des Atlas linguistiques régionaux. À cette époque, certains atlas n'étaient pas encore achevés et un grand nombre de données étaient restées inaccessibles. Aujourd'hui grâce au THESOC, certaines d'entres-elles sont devenues consultables (comme les données des vol. 3 de l'ALLOr et 4 de l'ALAL qui avaient été informatisées mais non publiées) ou le deviendront dans un proche avenir (il s'agit des données inédites de l'ALLOc dont il sera question ici, de l'ALLOr et de l'ALP).

[3] Cf. Brun-Trigaud, Darlu et *alii* (à paraître).

nouvelles notions lexicales[4] à saisir (soit un potentiel de 56.330 formes), j'ai imaginé d'utiliser les travaux dialectométriques de Camps (1986) sur l'ALLOc.

Suivant la méthode dite "globale" établie par Guiter (1973), dans laquelle chaque point est relié à ses plus proches voisins par le système classique de triangulation en usage dans cette discipline[5], Camps a procédé au comptage des différences pour chaque segment et établi un tableau complet des données[6] (1986, 125-131), la finalité de son étude étant de prouver l'existence de frontières dialectales à partir des résultats obtenus.

Dans ce tableau, servant de base, j'ai repéré le segment comportant le moins de différences pour chaque point d'enquête, ce qui a permis de construire un réseau reliant deux à deux les points, soit 130 segments. La carte 1[7], établie d'après les données de Camps, où l'épaisseur de la flèche qui relie les deux points est proportionnelle au nombre de différences brutes[8], montre bien que les difficultés allaient être plus grandes dans le nord du domaine ...

Ensuite, après avoir saisi l'intégralité des données du premier point (le choix s'est porté sur Toulouse (31.12) pour sa position centrale), j'ai continué en procédant par copier-coller et en ne modifiant que les formes différentes pour chacun des points, tout en suivant le réseau établi ci-dessus. Ce procédé a ainsi permis de saisir les 56.000 data en un temps relativement bref[9], ce qui montre que la dialectométrie peut également avoir d'autres usages que ceux que l'on connaît bien ...

---

[4] La morphologie nominale et la morphologie verbale ont déjà été saisies, elles seront jointes prochainement aux données du THESOC. Les champs lexicaux inédits se répartissent ainsi : corps humain (fin du chapitre), chronologie (jours de la semaine, etc.), "du berceau à la tombe" (jeux d'enfants, maladies, qualité, défauts, etc), la parenté, la religion et les fêtes religieuses, les métiers et la vie sociale.

[5] Cf. notamment Goebl (2012) pour une présentation détaillée.

[6] Hélas ce tableau n'est pas disponible dans l'étude menée pour l'ALLOr (Camps, 1991), seule la carte des résultats figure, ce qui ne me permettra peut-être pas de renouveler cette expérience pour les données inédites de cet atlas.

[7] Les cartes ont été établies à l'aide du logiciel Cartes & Données de Arctique et d'un sérieux coup de main de Photoshop ...

[8] Pour conserver une comparabilité des faits, seules les données brutes établies par Camps ont été utilisées.

D'autre part, je signale que j'ai parfaitement conscience de l'usage non conventionnel que je fais des cartes à rayons, puisque normalement ces dernières sont utilisées pour rendre compte du gradient de similarité entre deux points (cf. notamment Goebl 2012). Or je ne suis pas dans un réseau classique de dialectométrie avec triangulation, mais dans une relation de réseau deux-à-deux, d'où le choix de cette expression cartographique qui peut rendre compte à la fois de la similarité et de la distance par le jeu de l'épaisseur des traits reliant les points.

[9] Il n'a fallu que quatre mois, alors que normalement plus de six sont nécessaires pour la même quantité de saisie.

J'en ai profité pour éprouver ce que les méthodes dialectométriques pouvaient m'apporter sur ce domaine avec un corpus restreint.

2. Dans les pas de Guiter ...

Dans un premier temps, j'ai comparé les deux corpus : Camps avait travaillé sur un ensemble de 100 cartes comprenant le moins de lacunes possible, issues du premier volume de l'ALLOc[10], de même, j'ai sélectionné les 208 "cartes"[11] sans lacune de mon corpus.

Le même processus de comparaison que celui de Camps leur a été appliqué, c'est-à-dire en ne décomptant qu'une seule différence quelle que soit la nature (phonétique ou lexicale) ou le nombre de différences (dans le cas de variantes phonétiques d'un même mot) entre les deux segments[12] et en privilégiant la réponse la plus proche, dans le cas de réponses multiples. Le tableau en annexe (colonnes B, C et D) donne les résultats en pourcentage pour chacun des 130 segments dans les deux corpus.

Un constat apparaît immédiatement : globalement, mon corpus révèle des écarts plus grands entre les segments que ceux relevés par Camps (cf. le tableau en annexe, col. B et C). Comment expliquer cette disparité ? Guiter (1981) avait étudié cette question des écarts au sein sa propre méthode et en avait déduit que statistiquement un écart de 1% entre deux mesures (ce qu'il estimait pourtant comme très improbable) n'était pas significatif. C'est la proportion que nous trouvons ici, donc rien d'anormal[13].

3. ... Puis en suivant Séguy

Évidemment, on ne peut pas parler de dialectométrie sans évoquer J. Séguy qui créa le terme en 1973. Sa méthode était assez proche de celle de Guiter, à la différence qu'il prenait soin d'évaluer à part les variations phonologiques, phonétiques, morpho-syntaxiques et lexicales dont la moyenne était alors pondérée (cf. le vol. 6 de l'ALG).

---

[10] Camps, 1986, 123.

[11] L'usage du terme "carte" bien qu'impropre ici (aucune carte n'a été pour le moment créée avec les données) est destiné à rester dans l'univers familier des notions de la géolinguistique.

[12] En fait cette règle est tout à fait implicite chez Ch. Camps, comme chez H. Guiter : " [...] nous avons compté le nombre N de différences entre 2 points" (Camps, 1986, 120) ou "Le réseau de triangulation étant dressé, nous nous sommes proposé de rechercher combien de fois chaque segment était coupé par une ligne isoglosse" (Guiter, 1973, 67) ou "Entre chaque couple de points reliés, on compte le nombre de différences qui se manifestent, quelle que soit la nature de la (ou des) différence(s) constatée(s), d'où le nom de "méthode globale" donné à ce procédé extrêmement rapide." (Guiter, 1991, 101).

[13] En reprenant les données de Camps, avec l'aide du THESOC, il apparaît que le différentiel pourrait venir du fait que la distinction entre r: et ʀ n'a pas toujours été prise en compte.

J'ai moi-même séparé lexique et phonétique, afin d'établir leurs rôles respectifs dans l'appréhension des différences relevées entre les segments. Séguy a raison lorsqu'il affirme que : "Il n'en reste pas moins que le lexique est la clé de la compréhension : pour peu qu'on reconnaisse un certain nombre de mots dans une langue qu'on ne sait pas, on comprendra (relativement) l'énoncé ; inversement, si dans une langue qu'on manie, on rencontre un mot énigmatique, le message présente un trou que le contexte ne permet pas toujours de boucher." (1971, 339).

Les écarts lexicaux relevés sur les segments du réseau s'échelonnent de 2 à 21% avec une valeur moyenne de 10% (cf. le tableau en annexe, col. E et la carte 2). Les valeurs les plus importantes (en gris foncé dans le tableau et en noir sur la carte) ont été relevées essentiellement en périphérie, dans l'Ariège et dans l'Aude (ex. 09.10>09.01[14] : 16%), mais également dans le Lot, le Lot-et-Garonne, l'est du Tarn et en Dordogne (ex. 47.10>24.32 : 21%), qui de façon contrastée comporte aussi certaines des valeurs les plus faibles (en gris clair dans le tableau et sur la carte 2) (ex. 24.31>24.33 : 2%), que l'on retrouve dans l'ouest du Tarn (ex. 81.07>81.05 : 3%). Le lexique n'est donc pas réellement discriminant, sauf en périphérie du domaine, notamment à l'extrémité sud où l'on trouve un cumul important de valeurs relativement élevées.

Les écarts phonétiques se répartissent entre 10 et 78% avec une moyenne de 33% (cf. le tableau en annexe, col. F et la carte 3) si l'on continue de comptabiliser les faits avec la "méthode globale" de Guiter (différence = 1, quel que soit la nature ou le nombre de différences). Sans surprise, les plus faibles se rencontrent au centre et au sud du domaine (Tarn, Haute-Garonne, Tarn-et-Garonne et nord des départements de l'Ariège et de l'Aude) (en gris clair dans le tableau et sur la carte 3), mais les écarts les plus importants (en gris foncé sur le tableau et en noir sur la carte 3) se situent essentiellement au nord (Gironde, Lot-et-Garonne, Dordogne et Lot). Seule exception, le segment 11.22>09.33 situé dans l'espace en contact avec le catalan. Cette zone, plus fortement marquée du point de vue lexical comme nous venons de le voir, ne ressort donc pas particulièrement du point de vue phonétique.

La comparaison des cartes 2 et 3 montre qu'il existe, dans une certaine mesure, une corrélation entre les proportions des écarts lexicaux et phonétiques (faibles dans le centre du domaine, fortes dans le nord). Pourtant quelques segments font preuve de discordances (cf. 46.20>46.21 ou 82.04>82.03).

[14] La première localité citée est celle qui a servi de "modèle" pour la seconde.

4. Une nouvelle piste avec l'algorithme de Damerau-Levenshtein

Une analyse plus fine des faits phonétiques a été tentée en utilisant l'algorithme de Damerau-Levenshtein, mis en oeuvre, notamment, par l'école de dialectométrie de Groningen sous l'impulsion de J. Nerbonne et W. Heeringa[15].

Cet algorithme permet de mesurer la distance entre deux chaînes de caractères en se basant sur le nombre minimal de suppressions et/ou d'insertions et/ou de remplacement et/ou de transposition de caractères nécessaire pour passer d'une chaîne à l'autre. Nous y avons adjoint une nouvelle fonction[16] qui permet de retourner la valeur des caractères affectés par la mesure entre les deux chaînes comparées, et ainsi d'évaluer, pour chaque phonème de chaque segment, le nombre total d'opérations sur le corpus considéré. Il s'agit de déterminer si le score est élevé en raison d'un grand nombre de variations différentes ou par la répétition fréquente du même phénomène. L'intérêt de cette méthode, c'est qu'elle permet aussi d'analyser les mots ou les syntagmes dans leur globalité, tels qu'ils sont perçus par les locuteurs.

loc. source	chaîne source	opérations	nb	chaîne cible	loc. cible
46.14	bulũntsjˈɛ	− remplacement de u par ɔ − suppression de n − remplacement de ts par dz − suppression de j	4	bulɔ̃dzˈɛ	46.30
47.22	ʒurnˈaðɔ	− remplacement de r par ʁ − remplacement de ð par d − remplacement de ɔ par œ	3	ʒuʁnˈadœ	47.20

Tableau 1 : exemples de traitements avec l'algorithme de Damerau-Levenshtein

On se rend bien compte que si un phonème est récurrent, comme peut l'être /r/ ou encore la voyelle finale, alors sa variation lui confère un poids important dans les décomptes de la "méthode globale" ...

C'est ce qui a été observé de plus près pour chacun des 130 segments du réseau concernant les 208 "cartes" retenues, soit 27000 paires de chaînes

[15] Cf. notamment Heeringa, W. (2004) ou Nerbonne, J. & Heeringa, W. (2010). Tous leurs travaux sont disponibles sur leurs sites respectifs : http://urd.let.rug.nl/nerbonne/paper.html et http://urd.let.rug.nl/heeringa/dialectology/.

[16] En effet, les fonctions basées sur l'algorithme de Damerau-Levenshtein disponibles sur internet (cf. notamment la démo du site de Nerbonne et Heeringa, http://www.let.rug.nl/kleiweg/lev/) renvoient un score (le nombre d'opérations permettant de passer d'une chaîne à l'autre), mais pas la valeur des caractères modifiés. Nous avons utilisé la version de http://stackoverflow.com/questions/4243036/levenshtein-distance-in-excel et je remercie P.-A. Georges et S. Brun, informaticiens, pour leur précieuse aide.

analysées et soigneusement vérifiées. Globalement, environ 8000 chaînes ont été affectées par au moins une opération, soit 30% du corpus, pour un total d'environ 11000 (allant de une à six par chaînes comparées). Le tableau 2 regroupe les alternances les plus fréquentes :

alternances	nb	%	exemples
a/ɔ	612	5,59	atˈutʃ > ɔtˈutʃ (gifle)
r(:)/ʀ	594	5,42	ʃˈur > ʃˈuʀ (sourd)
ɛ/e	469	4,28	kɔrˈɛme > kɔrˈeme (carême)
d/ð	439	4,01	ʒurnˈadɔ > ʒurnˈaðɔ (journée)
s/ʃ	348	3,18	sˈɔw > ʃˈɔw (sou)
s (+/-)	299	2,73	yrˈys > yrˈy (heureux)
j (+/-)	297	2,71	gawtʃˈɛ > gawtʃjˈɛ (gaucher)
s/h	283	2,58	estˈiw > ehtˈiw (été)
t (+/-)	250	2,28	anˈɛjt > anˈɛj (ce soir)
nasalisation (+/-)	203	1,85	matˈĩŋ > matˈi (matin)
redoubl. cons. (+/-)	193	1,76	drˈɔle > drˈɔlːe (enfant)
r/rː	185	1,69	kˈɛre > kˈɛrːe (chercher)
n (+/-)	170	1,55	bulãdʒˈɛ > bulãndʒˈɛ (boulanger)
ts/dz	154	1,41	dutsˈenɔ > dudzˈenɔ (douzaine)
u/ɔ	151	1,38	prutestˈãnt > prɔtestˈãnt (protestant)
e/a	151	1,38	dʒãndˈarmɔ > dʒẽndˈarmɔ (gendarme)
ts/tʃ	128	1,17	mjɛtsˈũn > mjɛtʃˈũn (midi)
a/ɒ	147	1,34	brˈa > brˈɒ (bras)
e (+/-)	133	1,29	fasˈilːe > fasˈilː (facile)
b/β	137	1,25	djˈable > djˈaβle (diable)
g (+/-)	128	1,17	plˈagɔ > plˈaɔ (plaie)
ɔ/ɒ	126	1,15	pɔjrˈi > pɒjrˈi (parrain)
	5470	50 %	

Tableau 2 : alternances les plus fréquentes

On remarquera qu'en dépit de quantités relativement peu importantes, les 22 exemples ci-dessus regroupent la moitié du total enregistré sur les 400 alternances uniques relevées.
Qu'en est-il au niveau des segments ? Un inventaire des trois opérations les plus fréquentes pour chaque segment figure dans le tableau en annexe. Quelques faits intéressants en ont été extraits ici :

A	F	H	I	J	K	L	M	N
segment	diff. ph. %	opération 1	%	opération 2	%	opération 3	%	total %
2411>2410	55	R. de ð par d	7	R. de ɛ par e	7	R. de ʃ par ɕ	6	20
3112>3101	29	R. de l par w	6	S. de r	6	R. de ʎ par j	4	16
3132>8123	27	R. de ts par ʒ	32	R. de d par ð	19	R. de g par ɣ	6	57
4632>1202	78	R. de ʁ par r(:)	31	R. de tʃ par cʃ	7	Ins. de s	6	44
4710>2432	68	R. de s par ʃ	15	R. de ts par θ	13	R. de d par ð	11	39
4722>4720	71	R. de r(:) par ʁ	36	R. de ɔ par œ	22	R. de e par œ	10	68
8211>1203	39	R. de a par ɔ	32	R. de ^{s} par ^{h}	7	R. de s par ^{s}	6	45

Tableau 3 : exemples de répartition des opérations par segment

Deux possibilités se présentent pour expliquer une grande différence phonétique (col. F) : d'une part, dans le nord du domaine, la plupart des cas résultent d'une accumulation de variations (col. I, K, M ; gros rectangles + petites lignes grises, carte 4), cf. par ex. 47.10>24.32 ou 24.11>24.10 qui représente un des cas extrêmes d'éparpillement. D'autre part, mais un peu moins souvent, sur les marges orientale et occidentale de cette zone ainsi que pour le segment proche du domaine catalan (11.22>09.33), cette différence se manifeste par la prépondérance d'une seule variation (col. F, I ; gros rectangles grisés, carte 4), cf. 46.32>12.02 ou de deux, cf. 47.22>47.20.

Dans les segments à faibles différences phonétiques, on ne relève aucune valeur remarquable pour la première opération (col. F, I ; petits rectangles grisés, carte 4), cf. 31.12>31.01, mais, tout au contraire, plutôt une dispersion.

En revanche, dans les segments à valeurs moyennes, notamment dans le sud du Lot et du Lot-et-Garonne, ainsi que dans le Tarn, on relève la présence notable de certaines variations (col. I ; flèches noires dans aires grises, carte 4) cf. 31.32>81.23 ou 82.11>12.03.

Enfin, les phonèmes reportés sur la carte 5 montrent une forte présence des alternances r(:)/ʁ et a/ɔ dans le nord du domaine, comme le tableau 2 le laissait présager, tandis que l'est du Tarn est caractérisé par la présence de d/ð. On remarquera aussi que certains points sont le centre d'un nombre important de variations comme par ex. 46.14 ou 24.20 qui ainsi se singularisent de leurs voisins.

5. Conclusions

La saisie des données lexicales inédites de l'ALLOc m'a permis de tester, sur un nombre limité de paires de localités telles qu'elles ont été déterminées par le choix initial du corpus, une version personnelle de l'algorithme de Damerau-Levenshtein. Car, par-delà l'accumulation des chiffres tels que la statistique habituellement les engendre, il m'a semblé intéressant de pouvoir matérialiser ce qui constitue les obstacles à l'intercompréhension entre les différents segments étudiés.

Il serait instructif d'étendre cette recherche à tous les segments dans une triangulation complète, qui révèlerait encore davantage les clivages du domaine étudié[17].

Bibliographie

Boisgontier, Jacques, *Atlas linguistique et ethnographique du Languedoc oriental* (ALLOr), Paris, Ed. du CNRS, 1981-1986 (3 vol.).

Bouvier, Jean-Claude et Martel, Claude, *Atlas linguistique et ethnographique de la Provence* (ALP), Paris, Ed. du CNRS, 1975-1986 (3 vol.).

Brun-Trigaud, Guylaine, Darlu Pierre, Gaillard-Corvaglia Antonella, Léonard Jean Léo, Sauzet Patric, « Exploration cladistique de l'ALLOc », *Actes du Xe Congrès de l'AIEO (Béziers, 2011)*, à paraître.

Camps, Christian, « Limites linguistiques en Languedoc oriental », *Actes du XVIII^e Congrès International de Linguistique et de Philologie Romanes*, III, Tübingen, 1991, p. 362-69.

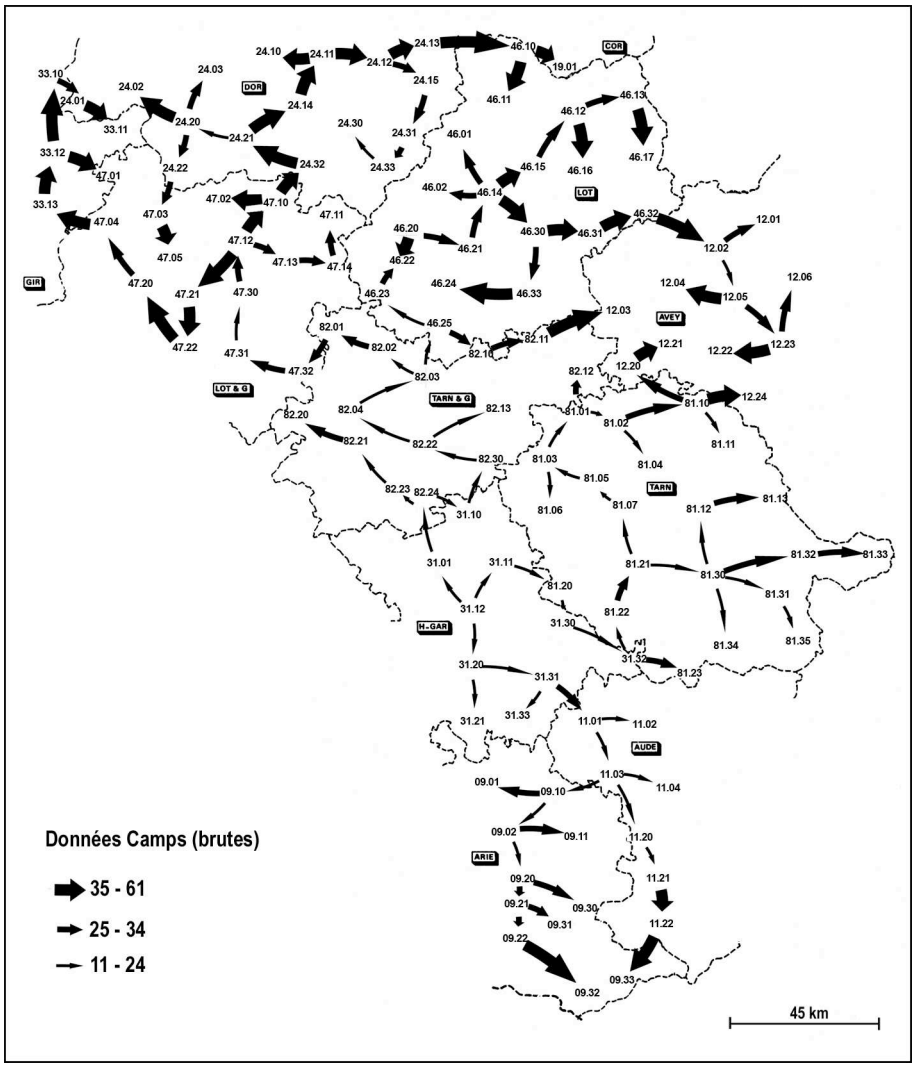
Camps, Christian, «Limites linguistiques d'après l'ALLOc », in *Variation linguistique dans l'espace : dialectologie et onomastique*. Actes du 17e Congrès International de Linguistique et de Philologie Romanes (Aix-en-Provence, 1983), Aix-en-Provence, Université de Provence, 1986, vol. 6, p. 117-135).

Dalbera, Jean-Philippe, *et al.*, *Thesaurus Occitan : 'THESOC'*, UMR 7320 BCL – CNRS / Université Nice Sophia Antipolis (1992-).

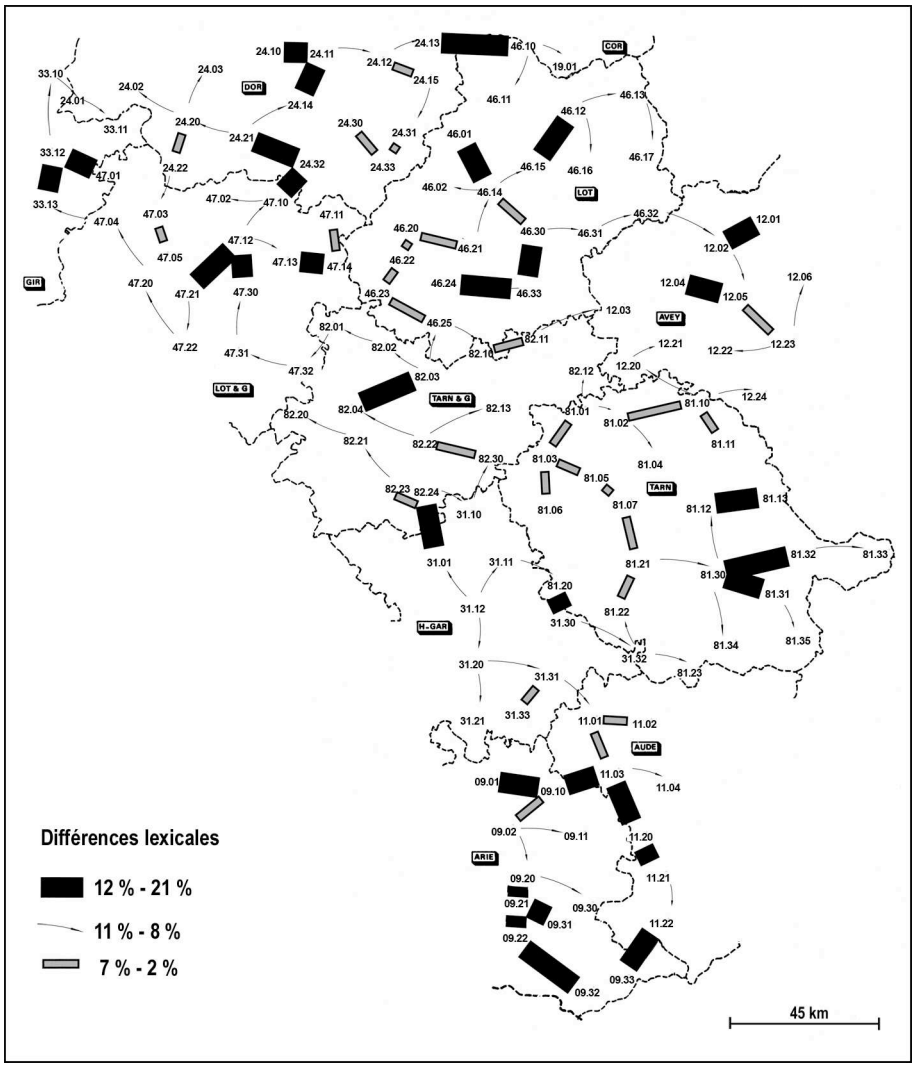
Goebl, Hans, « Introduction aux problèmes et méthodes de l'«École dialectométrique de Salzbourg» (avec des exemples gallo-, italo- et ibéroromans)», in: Álvarez Pérez, Afonso / Ernestina Carrilho / Catarina Magro (eds.) : *Proceedings of the International Symposium on Limits and Areas in Dialectology* (LimiAr, Lisbon 2011), Lisboa, Centro de Linguística da Universidade de Lisboa, 2012, p. 117-166.

[17] J'envisage aussi à brève échéance, avec le soutien de Hans Goebl, de tester également ces données dans le système de la DM de Salzbourg.

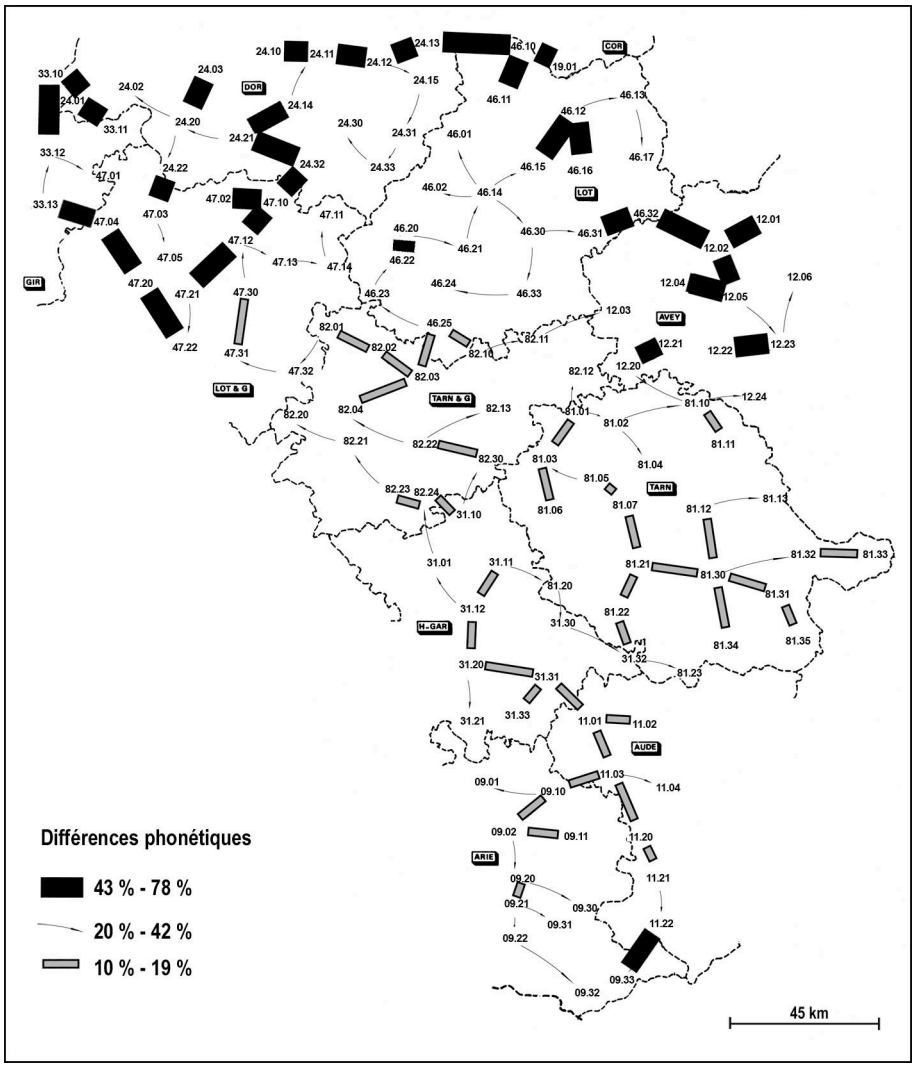
Guiter, Henri, « Appréciation de l'importance des écarts en géolinguistique », *Revue de Linguistique Romane*, 45, (1981), p. 341-348.
Guiter, Henri, « Atlas et frontières linguistiques », in *Les dialectes romans de la France à la lumière des atlas régionaux*, Paris, Ed. du CNRS, 1973, p. 61-109.
Guiter, Henri, « Sur l'*Atlas linguistique de l'Auvergne et du Limousin* », *Revue de Linguistique Romane*, 55, (1991), p. 100-117.
Heeringa, W, *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. Dissertation, University of Groningen, 2004.
Nerbonne, John et Heeringa, Wilbert, « Measuring dialect differences», in J.-E. Schmidt & P.Auer (eds.). *Language and Space: Theories and Methods*. Chap. 31. In series *Handbooks of Linguistics and Communication Science*, Berlin, Mouton de Gruyter, 2010, 550-567.
Potte, Jean-Claude, *Atlas linguistique et ethnographique de l'Auvergne et du Limousin* (ALAL), Paris, Ed. du CNRS, 1975-1992 (3 vol.).
Ravier, Xavier, *Atlas linguistique et ethnographique du Languedoc occidental* (ALLOc), Paris, Ed. du CNRS, 1978-1993 (4 vol.).
Séguy, Jean, « La relation entre la distance spaciale et la distance lexicale », *Revue de Linguistique Romane*, 35, (1971), p. 335-357.
Séguy, Jean, *Atlas linguistique et ethnographique de la Gascogne* (ALG), Paris, Ed. du CNRS, 1954-1974 (6 vol.).
Séguy, Jean, « La dialectométrie dans l'*Atlas linguistique de la Gascogne* », *Revue de Linguistique Romane*, 37, (1973), p. 1-24.



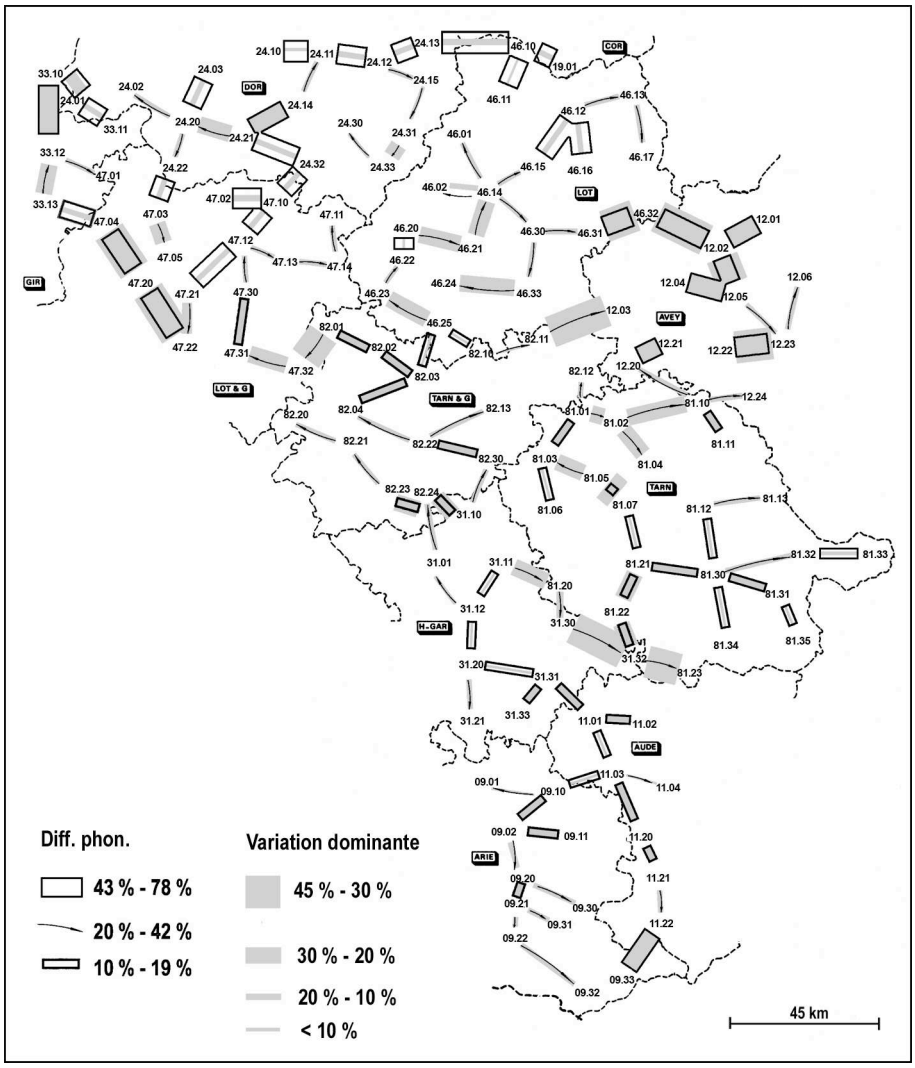
Carte 1 : réseau établi à partir des données brutes de Camps (1986)



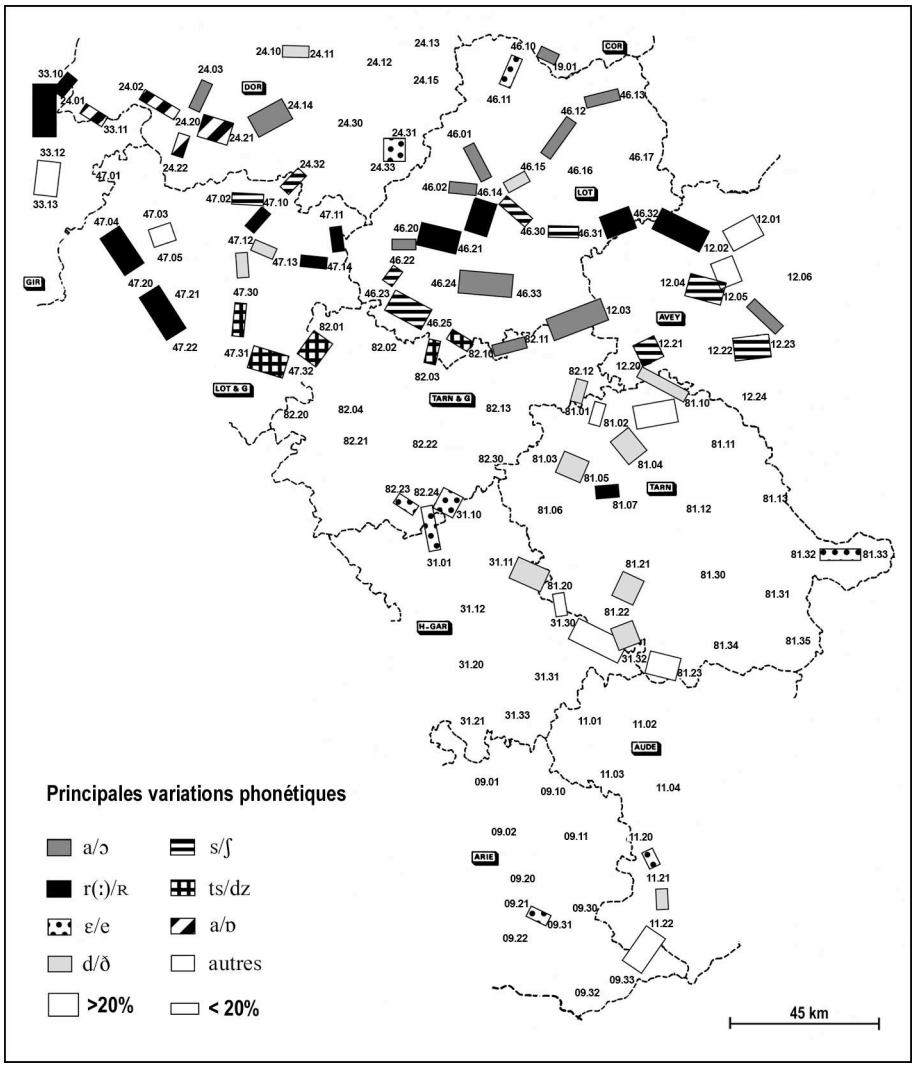
Carte 2 : Différences lexicales (valeurs maximales et minimales)



Carte 3 : Différences phonétiques (valeurs maximales et minimales)



Carte 4 : Différences phonétiques et proportion de
la variation dominante



Carte 5 : Principales variations phonétiques

ANNEXE : tableau récapitulatif

A : segments (Les chiffres entre crochets indiquent la moyenne)
B : C. Camps différences (brut) en % C: GBT différences en % D : Ecart entre Camps et GBT
E : GBT différences lexicale % F : GBT différences phonétiques % G : GBT nombre total de d'opérations
H : opération 1 et I : H / G J : opération 2 et K : J / G L : opération 3 et M : L /G N : I + K + M
Abréviations pour H, J, L : R. = remplacement ; S. = suppression ; Ins. = insersion
Cases gris clair : valeurs les plus faibles Cases gris foncé : valeurs les plus fortes

A	B [29]	C [39]	D	E [10]	F [33]	G [84]	H	I [17]	J	K [9]	L	M [7]	N [33]
segments	CC %	GBT %	écart	diff. lex. %	diff. ph. %	nb tot. chgmt	changement 1	%	changement 2	%	changement 3	%	total %
0902>0911	27	28	1	11	19	42	R. de dʒ par tʃ	10	R. de r par rː	10	R. de e par a	7	27
0902>0920	24	29	5	8	23	52	R. de ʎ par l int.	17	Ins. de s	6	R. de ʎ par l	7	30
0910>0901	26	34	8	16	21	44	R. de dʒ par ʒ	7	R. de tʃ par dʒ	7	R. de ʎ par l	5	19
0910>0902	22	21	-1	7	15	34	R. de tʃ par dʒ	12	Ins. de g	3	R. de ɔ par e	3	18
0920>0921	25	29	4	12	19	38	R. de l par l int.	11	R. de ts par dʒ	8	S. de s	8	27
0920>0930	32	32	0	9	25	72	R. de ʒ par ð	13	R. de ʒ par dʒ	6	S. de g	6	25
0921>0922	30	35	5	13	24	55	R. de l int. par l	11	R. de ð par d	9	R. de ɔ par o	7	27
0921>0931	30	37	7	13	27	54	R. de ɛ par e	17	R. de l int. par l	9	R. de ð par d	7	33
0922>0932	52	44	-8	13	34	78	R. de l par ʎ	14	R. de l it. par ʎ	6	S. de t	6	26
1101>1102	21	21	0	6	16	43	R. de y par œ	12	R. de dʒ par j	7	R. de tʃ par ts	7	26
1101>1103	19	17	-2	7	10	23	Ins. de t	9	R. de l par ʎ	9	R. de s par ʃ	9	27
1103>0910	20	27	7	13	16	32	Ins. de s	6	Ins. de t	6	R. de a par e	6	18
1103>1104	24	30	6	8	24	55	R. de ɔ par u	7	R. de d par ð	7	R. de ɔ par o	5	19
1103>1120	23	25	2	12	15	34	S. de la nasalis.	12	R. de l par ʎ	9	Ins. de g	6	27

A	B [29]	C [39]	D	E [10]	F [33]	G [84]	H	I [17]	J	K [9]	L	M [7]	N [33]
segments	CC %	GBT %	écart	diff. lex. %	diff. ph. %	nb tot. chgmt	changement 1	%	changement 2	%	changement 3	%	total %
1120>1121	22	23	1	12	13	27	R. de ɛ par e	11	R. de l par ʎ	11	R. de ɔ par a	7	29
1121>1122	35	36	1	11	28	63	R. de d par ð	13	R. de e par ɛ	10	Ins. de n	8	31
1122>0933	53	53	0	14	45	119	R. de ɔ par ə	29	R. de e par a	9	R. de ɔ par a	8	46
1202>1201	33	52	19	12	46	120	R. de cʃ par ts	26	R. de ɛ par e	7	R. de h par s	5	38
1202>1205	24	50	26	9	44	106	R. de cʃ par ts	31	R. de β par b	5	R. de r: par r	4	40
1205>1204	38	65	27	12	60	173	R. de s par ʃ	39	R. de ɔ par a	11	R. de b par β	6	46
1205>1223	32	38	6	7	33	79	R. de ɔ par a	11	R. de h par s	10	R. de j par ʎ	8	29
1220>1221	43	52	9	8	48	143	R. de s par ʃ	27	Ins. de cs. dble	14	S. de s	6	47
1223>1206	33	44	11	9	39	105	R. de ts par cʃ	13	R. de ts par ʒʲ	11	R. de a par ɔ	7	31
1223>1222	38	65	27	10	61	176	R. de s par ʃ	30	R. de ɔ par a	11	R. de ð par d	6	47
2401>3311	36	56	20	8	52	157	R. de a par ɒ	15	S. de j	13	R. de e par œ	8	36
2411>2410	41	61	20	14	55	149	R. de ð par d	7	R. de ɛ par e	7	R. de ʃ par ɕ	6	20
2411>2412	38	61	23	9	55	142	R. de ð par ts	17	R. de ʃ par ɕ	5	R. de θ par ts	5	27
2412>2413	41	62	21	8	58	166	R. de ɔ par ɒ	17	R. de ts par dz	13	R. de a par æ	5	35
2412>2415	32	43	11	4	41	107	R. de æ par a	11	R. de ɕ par ʃ	7	R. de ɛ par e	7	25
2413>4610	42	63	21	12	57	162	R. de ɒ par ɔ	19	R. de æ par a	9	R. de h par s	6	34
2414>2411	35	49	14	12	40	98	R. de a par æ	10	R. de ð par θ	9	R. de b par v	7	26
2415>2431	27	36	9	8	29	62	R. de ᵑ par n	6	R. de e par ɔ	5	R. de v par β	5	16
2420>2402	39	54	15	11	48	128	R. de ɒ par a	15	R. de ɕ par s	7	R. de ɛ par e	5	27
2420>2403	34	62	28	8	59	168	R. de a par ɔ	18	R. de ɒ par a	13	R. de ɕ par ʃ	8	39
2420>2422	29	45	16	6	42	107	R. de ɒ par a	12	R. de e par ɛ	6	S. de ŋ	6	24

A	B [29]	C [39]	D	E [10]	F [33]	G [84]	H	I [17]	J	K [9]	L	M [7]	N [33]
segments	CC %	GBT %	écart	diff. lex. %	diff. ph. %	nb tot. chgmt	changement 1	%	changement 2	%	changement 3	%	total %
2421>2414	42	50	8	11	44	115	R. de a par ɔ	29	R. de ɕ par ʃ	9	R. de θ par ð	6	44
2421>2420	22	44	22	8	39	98	R. de a par ɒ	20	R. de ʃ par ɕ	8	R. de ɛ par e	4	32
2422>4703	29	55	26	10	50	123	R. de ɕ par s	17	R. de ɒ par a	9	R. de e par ɛ	7	33
2431>2433	26	30	4	2	28	72	R. de e par ɛ	25	R. de ɔ par a	7	R. de ʃ par ɕ	7	39
2432>2421	38	55	17	12	49	122	R. de θ par ð	17	R. de ʃ par ɕ	7	R. de β par v	6	30
2433>2430	23	38	15	3	35	97	R. de ʃ par ɕ	9	R. de ɛ par e	8	R. de ɔ par a	6	23
3101>8224	20	24	4	12	26	61	R. de ɛ par e	17	R. de w par l	8	Ins. de s	5	30
3110>8230	19	33	14	10	25	49	S. de j	16	Ins. de ʃ	7	R. de ts par tʃ	5	28
3111>8120	20	31	11	10	23	63	R. de d par ð	24	R. de tʃ par c	10	R. de dʒ par ʒ	4	38
3112>3101	17	36	19	10	29	69	R. de l par w	6	S. de r	6	R. de ʎ par j	4	16
3112>3111	15	22	7	10	13	29	Ins. de s	7	R. de ð par d	7	R. de r par rː	7	21
3112>3120	14	25	11	10	16	40	Ins. de t	8	R. de ð par d	5	Ins. de g	5	18
3120>3121	21	30	9	8	21	46	R. de l par ʎ	13	R. de d par ð	7	R. de g par ɣ	7	27
3120>3131	19	23	4	10	14	35	R. de dʒ par j	6	R. de e par u	6	R. de r par rː	6	18
3130>3132	24	35	11	10	28	64	R. de ɟ par ts	30	R. de c par ts	9	R. de ʒ par ts	6	45
3131>1101	25	24	-1	9	17	45	R. de l par ʎ	11	R. de ts par tʃ	7	S. de t	4	22
3131>3133	19	21	2	7	14	36	R. de l par ʎ	14	Ins. de j	11	R. de n par ɲ	6	31
3132>8122	21	25	4	8	19	47	R. de d par ð	23	Ins. de s	11	S. de j	11	45
3132>8123	26	34	8	10	27	62	R. de ts par ʒ	32	R. de d par ð	19	R. de g par ɣ	6	57
3310>2401	31	64	33	9	60	169	R. de ʀ par r(ː)	25	R. de œ par e	18	R. de ɔ par u	5	48
3312>3310	37	62	25	8	58	167	R. de r(ː) par ʀ	25	R. de u par ɔ	5	R. de b par v	4	34

A	B [29]	C [39]	D	E [10]	F [33]	G [84]	H	I [17]	J	K [9]	L	M [7]	N [33]
segments	CC %	GBT %	écart	diff. lex. %	diff. ph. %	nb tot. chgmt	changement 1	%	changement 2	%	changement 3	%	total %
3312>4701	39	41	2	12	33	79	R. de u par ɔ	10	R. de ɛ par e	9	R. de j par ʎ	9	28
3313>3312	37	48	11	15	38	100	S. de t	22	R. de ɛ par e	6	R. de a par ɔ	4	32
4610>1901	54	67	13	9	61	179	R. de ɔ par a	13	R. de ɔ par ɒ	13	R. de d par ð	8	34
4610>4611	41	55	14	9	50	136	R. de e par ɛ	8	R. de a par ɔ	6	R. de v par β	6	20
4612>4613	28	43	15	9	37	95	R. de a par ɔ	17	R. de ð par d	4	R. de rː par r	4	25
4612>4616	38	58	20	8	55	144	R. de ts par tʃ	18	Ins. de w	8	R. de a par ɔ	8	34
4613>4617	40	42	2	9	37	97	Ins. de t	16	R. de a par ɔ	11	Ins. de w	10	37
4614>4601	27	40	13	13	32	77	R. de ɔ par a	10	R. de ts par dz	6	S. de t	5	21
4614>4602	28	33	5	10	26	66	R. de ɔ par a	12	R. de ʃ par s	11	R. de ʒ par z	8	31
4614>4615	36	46	10	10	40	96	R. de d par ð	17	R. de ^{h} par s	10	R. de ʃ par ɕ	7	34
4614>4630	39	43	4	6	39	106	R. de ʃ par s	13	R. de ^{h} par s	9	R. de ts par dz	7	29
4615>4612	33	55	22	12	49	143	R. de ɔ par a	14	R. de ð par d	8	R. de n par ŋ	8	30
4620>4621	29	50	21	7	46	118	R. de r(ː) par ʁ	24	R. de ɔ par a	7	R. de ʃ par s	7	38
4621>4614	32	52	20	8	48	108	R. de ʁ par r(ː)	24	S. de t	10	R. de s par ʃ	7	41
4622>4620	41	50	9	4	48	139	R. de a par ɔ	45	S. de t	6	S. de ʃ	5	56
4623>4622	28	39	11	7	34	79	R. de s par ʃ	16	R. de s par ^{h}	14	S. de s	14	44
4625>4623	23	38	15	7	33	78	R. de s par ʃ	28	S. de s	9	R. de dz par ts	4	41
4625>8210	25	23	-2	9	16	35	R. de dz par ts	9	R. de h par ^{h}	9	S. de j	6	24
4630>4631	38	37	-1	8	31	80	R. de s par ʃ	16	Ins. de n	13	R. de ɔ par a	9	38
4630>4633	34	49	15	16	39	85	Ins. de s	14	R. de ʃ par s	14	R. de ɔ par a	7	35
4631>4632	47	76	29	11	74	205	R. de r(ː) par ʁ	37	R. de ʃ par s	8	R. de ts par tʃ	6	51

A	B [29]	C [39]	D	E [10]	F [33]	G [84]	H	I [17]	J	K [9]	L	M [7]	N [33]
segments	CC %	GBT %	écart	diff. lex. %	diff. ph. %	nb tot. chgmt	changement 1	%	changement 2	%	changement 3	%	total %
4632>1202	47	80	33	10	78	235	R. de ʁ par r(ː)	31	R. de tʃ par cʃ	7	Ins. de s	6	44
4633>4624	35	49	14	18	37	87	R. de ɔ par a	22	R. de s par ʃ	16	R. de s par ^{h}	8	46
4703>4705	36	36	0	6	31	72	R. de ð par ʒ	28	R. de ^{m} par ŋ	11	R. de e par ɛ	10	49
4704>3313	45	50	5	11	44	122	R. de e par œ	18	R. de ʁ par rː	10	R. de e par ə	6	34
4710>2432	43	75	32	21	68	172	R. de s par ʃ	15	R. de ts par θ	13	R. de d par ð	11	39
4710>4702	35	62	27	8	58	143	R. de s par ʃ	15	R. de ts par dz	15	R. de ʁ par rː	13	43
4712>4710	38	63	25	9	59	151	R. de rː par ʁ	15	R. de ɛ par e	10	S. de t	9	34
4712>4713	26	42	16	11	35	82	R. de ð par d	12	R. de ɛ par e	5	Ins. de s	4	21
4712>4721	36	58	22	13	52	116	R. de ts par ʒ	16	R. de rː par ʁ	9	R. de ^{h} par s	5	30
4713>4714	28	36	8	12	27	58	R. de r par ʁ	14	R. de ɛ par e	12	S. de n	10	36
4714>4711	27	39	12	7	35	84	R. de rː par ʁ	18	S. de s	13	R. de s par ^{h}	11	42
4720>4704	33	65	32	10	61	166	R. de ʁ par r	39	R. de d par ð	11	Ins. de j	5	55
4721>4722	36	42	6	9	36	101	Ins. de ŋ	12	R. de n par ŋ	12	Ins. de nasalis.	10	34
4722>4720	61	74	13	11	71	206	R. de r(ː) par ʁ	36	R. de ɔ par œ	22	R. de e par œ	10	68
4730>4712	25	47	22	14	38	87	R. de d par ð	14	R. de t par ^{t}	10	R. de s par ^{s}	9	33
4731>4730	22	27	5	10	19	47	R. de dz par ts	11	R. de ʃ par ts	6	R. de e par a	4	21
4732>4731	27	35	8	11	27	57	R. de dz par ts	28	R. de h par ^{h}	11	R. de r par rː	5	44
8101>8102	24	29	5	11	20	45	R. de l par r	26	R. de a par e	4	R. de b par β	4	35
8101>8212	32	39	7	10	32	72	R. de ð par d	19	R. de a par ɔ	13	S. de j	7	39
8102>8104	24	31	7	9	23	53	R. de ð par d	25	R. de r par l	17	R. de ð par r	6	48
8102>8110	28	26	-2	6	21	54	R. de s par ^{h}	26	Ins. de e	4	R. de l par r	4	34

A	B [29]	C [39]	D	E [10]	F [33]	G [84]	H	I [17]	J	K [9]	L	M [7]	N [33]
segments	CC %	GBT %	écart	diff. lex. %	diff. ph. %	nb tot. chgmt	changement 1	%	changement 2	%	changement 3	%	total %
8103>8101	15	19	4	7	13	28	S. de t	11	R. de g par ɣ	7	R. de r par rː	7	25
8103>8106	23	23	0	7	17	37	Ins. de ɣ	8	R. de ð par r	8	S. de t	8	24
8105>8103	21	33	12	5	29	65	R. de d par ð	25	R. de ʁ par r(ː)	23	R. de g par ɣ	6	54
8107>8105	18	19	1	3	16	40	R. de rː par ʁ	38	R. de d par r	8	R. de b par β	5	51
8110>1220	34	41	7	9	35	93	R. de ð par d	16	R. de ɣ par g	6	R. de ^{h} par s	5	27
8110>1224	38	41	3	9	35	90	R. de ^{h} par s	12	R. de r par l	12	R. de ð par d	6	30
8110>8111	22	21	-1	7	14	35	R. de ^{h} par s	14	R. de r par l	6	R. de rː par r	6	26
8112>8113	28	48	20	12	41	95	R. de s par ^{h}	15	R. de d par ð	13	R. de l par r	12	40
8120>3130	20	35	15	12	26	59	R. de ʒ par ɟ	24	R. de ð par d	20	R. de s par h	5	49
8121>8107	12	19	7	5	14	35	Ins. de s	9	S. de j	9	Ins. de g	6	24
8121>8130	16	20	4	10	10	19	Ins. de e	11	Ins. de s	11	Ins. de g	5	27
8122>8121	25	22	-3	5	17	42	R. de ð par d	26	R. de e par a	5	Ins. de g	5	36
8130>8112	21	23	2	11	13	27	R. de r par rː	7	R. de s par h	7	S. de ː	7	21
8130>8131	20	24	4	13	13	31	S. de cs. dble	10	Ins. de g	6	R. de β par b	6	22
8130>8132	28	32	4	12	22	57	R. de l par r	14	R. de l par j	12	Ins. de s	7	33
8130>8134	15	21	6	8	14	35	S. de s	9	R. de g par ɣ	6	R. de i par e	6	21
8131>8135	23	20	-3	11	10	23	S. de s	9	Ins. de ɣ	4	R. de d par ð	4	17
8132>8133	25	27	2	10	18	49	R. de e par ɛ	8	S. de nasalis.	8	R. de u par ɔ	4	20
8201>4732	27	35	8	9	28	67	R. de ts par dz	37	R. de ^{h} par h	7	R. de s par ^{s}	6	50
8202>8201	27	25	-2	9	18	44	S. de s	11	R. de ^{s} par s	9	R. de e par ɛ	5	25
8203>4625	21	25	4	8	18	40	R. de ts par dz	8	R. de ^{h} par h	10	S. de k	8	26

A	B [29]	C [39]	D	E [10]	F [33]	G [84]	H	I [17]	J	K [9]	L	M [7]	N [33]
segments	CC %	GBT %	écart	diff. lex. %	diff. ph. %	nb tot. chgmt	changement 1	%	changement 2	%	changement 3	%	total %
8203>8202	20	26	6	9	19	47	S. de nasalis.	15	R. de s par ^{s}	9	Ins. de ɡ	6	30
8204>8203	20	30	10	13	19	45	R. de s par ^{h}	16	Ins. de nasalis.	11	R. de r: par r	9	36
8210>8211	25	31	6	6	27	73	R. de a par ɔ	8	R. de a par ɔ	8	R. de ^{h} par ^{s}	8	24
8211>1203	42	46	4	11	39	98	R. de a par ɔ	32	R. de ^{s} par ^{h}	7	R. de s par ^{s}	6	45
8221>8220	29	42	13	10	35	91	Ins. de nasalis.	10	R. de r: par ʀ	10	Ins. de n	8	28
8222>8204	23	35	12	11	26	71	R. de tʃ par ts	13	R. de ʒ par ts	13	R. de dʒ par ts	10	36
8222>8213	20	32	12	10	24	58	R. de tʃ par ts	16	R. de ʒ par ts	16	R. de dʒ par ts	10	42
8223>8221	17	29	12	8	23	58	S. de j	10	S. de nasalis.	7	S. de t	7	24
8224>3110	17	35	18	8	18	41	R. de e par ɛ	22	Ins. de ɡ	5	S. de s	5	32
8224>8223	11	25	14	6	19	47	R. de e par ɛ	23	R. de r par r:	9	R. de e par a	4	36
8230>8222	18	18	0	5	14	37	R. de tʃ par ts	14	R. de tʃ par dʒ	8	R. de ʒ par ts	8	30