



**HAL**  
open science

## An unforeseen use of the Damerau-Levenshtein algorithm in the Occitan area

Guylaine Brun-Trigaud

► **To cite this version:**

Guylaine Brun-Trigaud. An unforeseen use of the Damerau-Levenshtein algorithm in the Occitan area. 2014. halshs-01067300

**HAL Id: halshs-01067300**

**<https://shs.hal.science/halshs-01067300v1>**

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.

**Guylaine BRUN-TRIGAUD**

IE1 CNRS Laboratoire Bases-Corpus-Langage (UMR 7320)

Université de Nice Sophia Antipolis (France)

Texte original : "Un usage particulier de l'algorithme de Damereau-Levenshtein dans le domaine occitan", in *Toujours langue varie ... Mélanges Andres Kristol*, édités par F. Diemoz et D. Aquino-Weber, Droz, Genève, 2014, 127-147.

## **An unforeseen use of the Damerau-Levenshtein algorithm in the Occitan area**

### **1 Introduction**

When I had finished entering into our basis the data of the last maps of the regional linguistic atlases published by the CNRS, which constitute the main source of the Thesaurus Occitan (THESOC) (nearly 800,000 data are currently available online<sup>1</sup>), the scientific project leaders, Jean-Philippe Dalbera and Michele Olivieri entrusted me with the task of completing this corpus with unpublished data from Alloc and Allor<sup>2</sup>.

Until then the THESOC data had always been collected in published maps, which allowed me a frequent use of the copy and paste functions, especially in a relatively homogeneous area like Alloc<sup>3</sup>. But this solution seemed no longer possible since the data were to be picked up directly from the survey notebooks (it is impossible to check at the same time the answers for the same concept from each of the 131 notebooks, even in the form of scanned images,...).

So as to save time without embarking on the chore of taking one by one, for each of the 131 survey items, the 430 new lexical concepts<sup>4</sup> to input (i.e. a potential 56,330 forms), I had the idea to apply Camps's dialectometric method (1986) to Alloc.

Following the so-called "global" method established by Guiter (1973), in which each point is connected to its nearest neighbours by the conventional triangulation system used in this discipline<sup>5</sup>, Camps proceeded to count the differences for each segment and built a full table of data<sup>6</sup> (1986, p. 125-131). The purpose of his study was to prove the existence of dialect boundaries based on the results achieved.

Using that table as a basis, I spotted the segment which showed the smallest number of differences for each survey point, which helped me build a network connecting the points two

---

<sup>1</sup> <http://thesaurus.unice.fr/>.

<sup>22</sup>. In 1996 CNRS put an end to the publication of the Regional Linguistic Atlas of France. At that time, some atlases had not yet been completed and a large number of data remained inaccessible. Now thanks to THESOC some of them are searchable (such as data from Allor 3 and ALAL 4 which had been computerized but not published) or will be accessible in the near future (they consist of unpublished data from Alloc which will be discussed here, from Allor and from ALP).

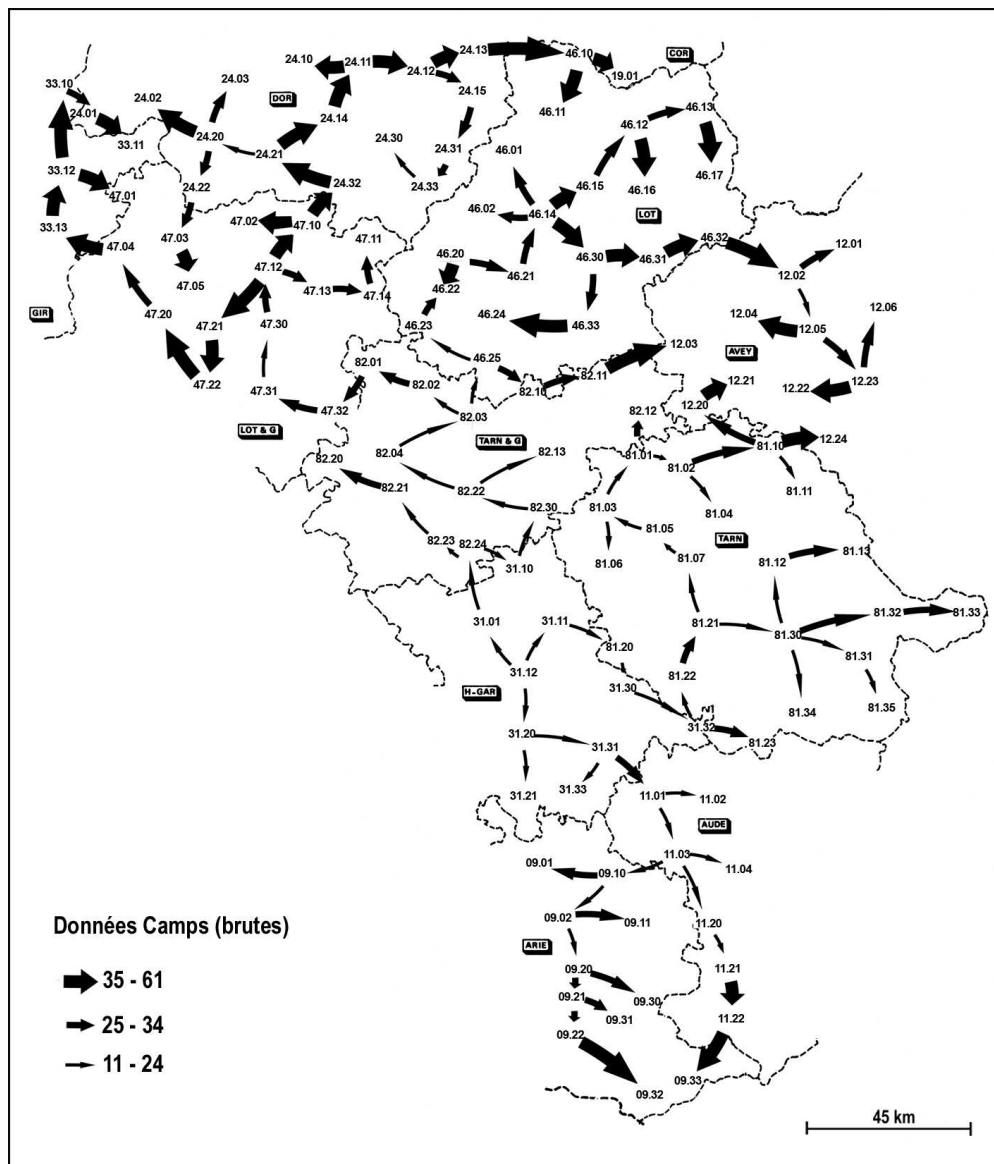
<sup>3</sup> Cf. Brun-Trigaud, Darlu et alii (forthcoming).

<sup>4</sup> Nominal and verbal morphologies have already been entered and will be added shortly to the THESOC data. The unpublished lexical fields are distributed as follows: human body (end of chapter), timeline (days of the week, etc.), "from cradle to grave" (children's games, diseases, qualities, defects, etc.), kinship, religion and religious festivals, trades and social life

<sup>5</sup> See in particular Goebel (2012) for a detailed presentation.

<sup>66</sup> This table is unfortunately not available in the study conducted for Allor (Camps, 1991). Only the results map is found in it, which means that perhaps I will not be able to repeat this experience for unpublished data from this atlas.

by two, in all 130 segments. Map 1<sup>7</sup>, drawn starting from Camps's data, where the thickness of the arrow connecting the two points is proportional to the gross differences<sup>8</sup>, shows clearly that the difficulties would be greater in the northern part of the area ...



Map 1: network established from the raw data Camps (1986)

After I had entered all the data from the first point (the choice fell on Toulouse (31.12) because of its central position), the next step was to continue as before using the copy and paste function, changing nothing but the items that were different for each point, while following the network established above. I was thus able to enter 56,000 data in a relatively

<sup>7</sup> The maps were produced using the software Cartes & Données de Arctique and a good help by Photoshop ...

<sup>8</sup> To maintain the comparability of the facts, only the raw data set by Camps were used.

On the other hand, I acknowledge that I am aware of my unconventional use of rays cards, since these are normally used to account for the gradient of similarity between two points (see in particular Goebel 2012). But I do not work within a conventional dialectometry with triangulation network, but in a two-to-two relationship, hence the choice of this cartographic expression that can account for both similarity and distance set in the thickness of the lines connecting the points.

short time<sup>9</sup>, which shows that dialectometry may also be put to serve purposes different from those we are familiar with...

I took that opportunity to experience how dialectometric methods could help me with my work on that area starting from a small corpus.

## 2 Following Guiter's steps ...

To start with, I compared the two corpora: Camps had been working on a set of 100 maps with minimal gaps from the first volume of *Alloc*<sup>10</sup>. I likewise selected the 208 complete "maps"<sup>11</sup> (i.e. with no gaps) available in my corpus.

I applied to them the same comparison method Camps had used, that is to say, counting only one difference whatever its nature (either lexical or phonetic) or the number of phonetic variants between the two segments<sup>12</sup>, focusing on the nearest form, in the case of multiple responses. The attached table (columns B, C and D) shows the percentage results for each of the 130 segments in both corpora.

One is immediately struck by the fact that the number of differences between segments is larger in my corpus than in Camps's (see attached table, col B and C.). What accounts for this disparity? Guiter (1981) had studied the issue of differences within his own method and deduced that statistically a 1% difference between two measurements (what he considered to be highly unlikely) was not significant. We find the same proportion here, which means that is nothing went amiss<sup>13</sup>.

## 3 ... Then following Séguy's

One cannot obviously talk about dialectometry without mentioning J. Séguy who coined the term in 1973. His method was quite similar to Guiter's, except that he decided to set apart phonological, phonetic, morphosyntactic and lexical variations whose average was then weighted (see vol. 6 ALG).

I have myself treated apart phonetic and lexical variations so as to determine their respective roles in the understanding of the differences between the segments. Séguy rightly states: "The fact remains that the lexicon is the key to comprehension: we only have to recognize a number of words in a language we do not know, to understand what a statement is about; conversely, if in a language we know we encounter an enigmatic word, this leaves a gap in the message that the context does not always allow us to fill." (1971, 339).

The lexical gaps identified on the segments of the network range from 2-21% with an average value of 10% (see the attached table, col. E and Map 2). The most important values (dark gray

---

<sup>99</sup> It took me only four months when normally more than six are required for the same amount of input.

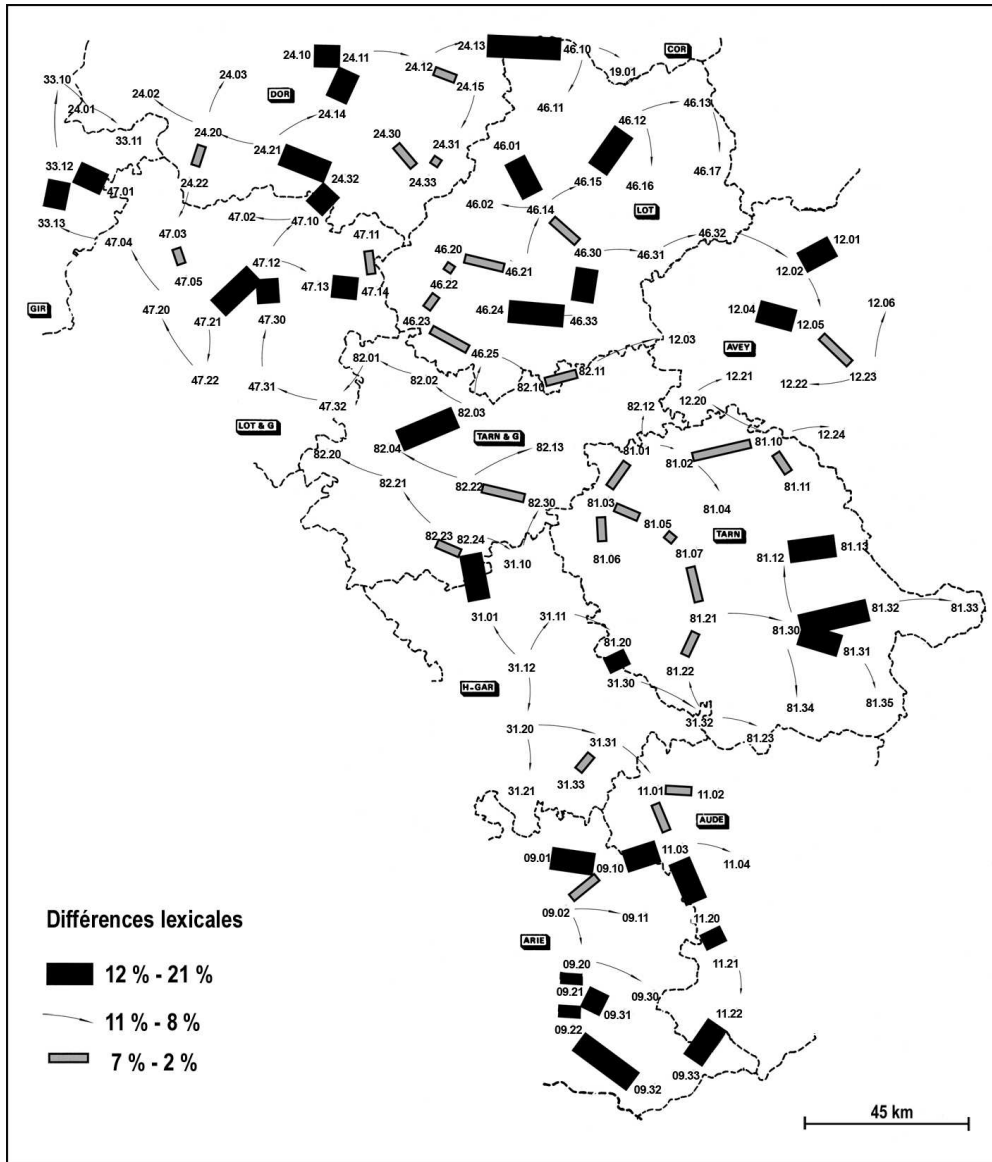
<sup>1010</sup> Camps, 1986, 123.

<sup>1111</sup> The use of the word "map" although unfit here (no map was currently created with the data) is intended to remain in the familiar world of geolinguistic concepts.

<sup>1212</sup> In fact this rule is quite implicit in Ch Camps, as in H. Guiter. "[...] We counted the number N of differences between 2 points" (Camps 1986, 120) or "network triangulation having been prepared, we have offered to investigate how often each segment was cut by a isogloss line "(Guiter, 1973, 67) or" between each pair of connected points, we count the number of differences that occur, regardless of the nature of the difference(s) found. Hence the name "comprehensive approach" given to this extremely fast process." (Guiter, 1991, 101).

<sup>13</sup> Using Camps's data, with the help of THESOC, it appears that the difference could be that the distinction between R and □ has not always been taken into account.

in the table and black on the map) were mainly found in the periphery, in Ariège, Aude (ex 09.10 > 09.01<sup>14</sup>: 16%) but also in Lot, Lot-et-Garonne, Tarn and eastern Dordogne (ex 47.10 > 24.32: 21%), which paradoxically also have some of the lowest values (light gray in the table and on map 2) (ex 24.31 > 24.33: 2%), which are also found in western Tarn (ex 81.07 > 81.05: 3%). Thus the lexicon is not really discriminating, except in the periphery, particularly at the southern end where there is a significant accumulation of relatively high values.



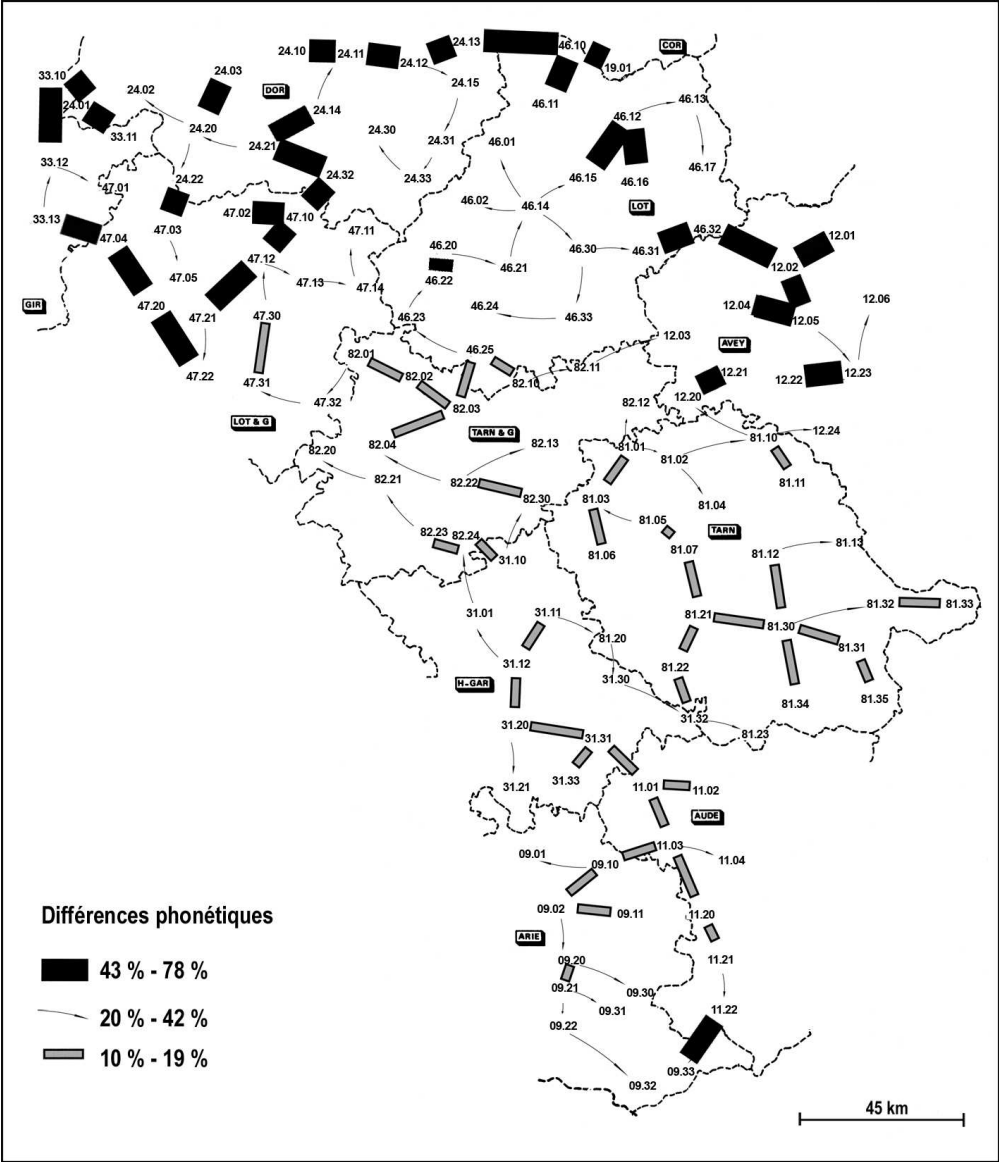
Map 2: lexical differences (maximum and minimum values)

Phonetic differences fall between 10 and 78% with an average of 33% (see the attached table, col. F and map 3) if we continue to use Guiter's "global method" (difference = 1, regardless of the nature or number of differences). Not surprisingly, the lowest are found in the central and southern areas (Tarn, Haute-Garonne, Tarn-et-Garonne and the northern departments of Ariège and Aude) (light gray in the table and on map 3), but the largest differences (in dark gray in the table and in black on map 3) lie mainly in the North (Gironde, Lot-et-Garonne, Lot and Dordogne). The only exception is segment 11.22 > 09.33 in contact with the Catalan-

<sup>1414</sup> The first location quoted served as a "model" for the second.

speaking zone. This area, lexically more marked as we have just seen, does not particularly stand out from a phonetic viewpoint.

A comparison between maps 2 and 3 shows that there is a certain degree of correlation between the proportion of lexical and phonetic differences (low in the centre of the field, strong in the North). Yet some segments show some discrepancies (see 46.20 > 46.21 or 82.04 > 82.03).



Map 3: phonetic differences (maximum and minimum values)

#### 4 A new track with the Damerau-Levenshtein algorithm

I have attempted a more detailed analysis of the phonetic facts using the Damerau-Levenshtein algorithm, which was notably introduced by the Groningen dialectometry school under the leadership of J. Nerbonne and W. Heeringa<sup>15</sup>.

This algorithm allows to measure the distance between two strings based on the minimum number of deletions and/or insertions and/or replacements and/or character transpositions involved to switch from one string to another. We have added a new function<sup>16</sup> allowing to return the value of the characters affected by the measurement between the two compared strings, and so to evaluate, for each phoneme of each segment, the total number of operations required on the corpus under consideration. This is to determine whether the score is high because of the number of different variations or because of the frequent repetition of the same phenomenon. The appeal of this method is that it also allows to analyze the words or phrases in their entirety, as they are heard by the speakers.

loc. source	source string	operations	nb	target string	target loc.
46.14	bulũntsʲɛ	- replacement of u by ɔ - suppression de n - replacement of ts by dz - removal of j	4	bulɔdz'ɛ	46.30
47.22	ʒurn'aðɔ	- replacement of r by R - replacement of ð by d - replacement of ɔ by œ	3	ʒurn'adœ	47.20

Tableau 1 : examples of treatments using Damerau-Levenshtein algorithm

It is clear that if a phoneme like / r / or the final vowel is recurring, then its variation gives it a substantial weight within the totals of the "global method" ...

This is what we have closely noted for each of the 130 segments of the network on the 208 "maps" retained, viz 27,000 pairs of analyzed and carefully checked strings. Overall, about 8,000 strings were affected by at least one operation, i.e. 30% of the corpus, for a total of about 11,000 (ranging from one to six in each of the strings under comparison). Table 2 summarizes the most frequent alternations:

<sup>15</sup> See in particular Heeringa, W. (2004) and Nerbonne, J. & Heeringa, W. (2010). All their works are available on their respective websites: <http://urd.let.rug.nl/nerbonne/paper.html> and <http://urd.let.rug.nl/heeringa/dialectology/>.

<sup>16</sup> In fact, the functions based on the Damerau-Levenshtein algorithm available online (see in particular the Nerbonne and Heeringa demo site, <http://www.let.rug.nl/kleiweg/lev/>) return a score (number of operations necessary to move from one string to another), but not the value of the modified characters. We used the version of <http://stackoverflow.com/questions/4243036/levenshtein-distance-in-excel,-and> I thank P. A. Georges and S. Brun, computer scientists, for their invaluable assistance.

alternations	nb	%	examples
a/ɔ	612	5,59	at'utʃ > ɔt'utʃ (slap)
r(:)/ʀ	594	5,42	ʃ'ur > ʃ'uʀ (deaf)
ɛ/e	469	4,28	kɔr'ɛme > kɔr'eme (Lent)
d/ð	439	4,01	ʒurn'adɔ > ʒurn'aðɔ (day)
s/ʃ	348	3,18	s'ɔw > ʃ'ɔw (penny)
s (+/-)	299	2,73	yr'ys > yr'y (happy)
j (+/-)	297	2,71	gawtʃ'ɛ > gawtʃj'ɛ (left-handed)
s/h	283	2,58	est'iw > eht'iw (summer)
t (+/-)	250	2,28	an'ejt > an'ej (tonight)
nasalisation (+/-)	203	1,85	mat'ĩŋ > mat'i (morning)
redoubl. cons. (+/-)	193	1,76	dr'ɔle > dr'ɔlɛ (child)
r/r:	185	1,69	k'ere > k'ɛrɛ (to seek)
n (+/-)	170	1,55	bulãdz'ɛ > bulãndz'ɛ (baker)
ts/dz	154	1,41	duts'enɔ > dudz'enɔ (dozen)
u/ɔ	151	1,38	prutest'ãnt > prɔtest'ãnt (Protestant)
e/a	151	1,38	dʒãnd'armɔ > dʒẽnd'armɔ (gendarme)
ts/tʃ	128	1,17	mjets'ũn > mjetʃ'ũn (noon)
a/ɔ	147	1,34	br'a > br'ɔ (arm)
e (+/-)	133	1,29	fas'il:ɛ > fas'il: (easy)
b/β	137	1,25	dj'able > dj'aβle (devil)
g (+/-)	128	1,17	pl'agɔ > pl'aɔ (wound)
ɔ/ɔ	126	1,15	pɔjr'i > pɔjr'i (godfather)
	5470	50 %	

Table 2: Most frequent alternations

Note that, despite the relatively small numbers, the above 22 examples include half of the 400 recorded total unique variations detected.

What happens at segment level? An inventory of the three most common operations for each segment is provided in the attached table. Some interesting facts have been extracted here:

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

Table 3: Sample breakdown by segment

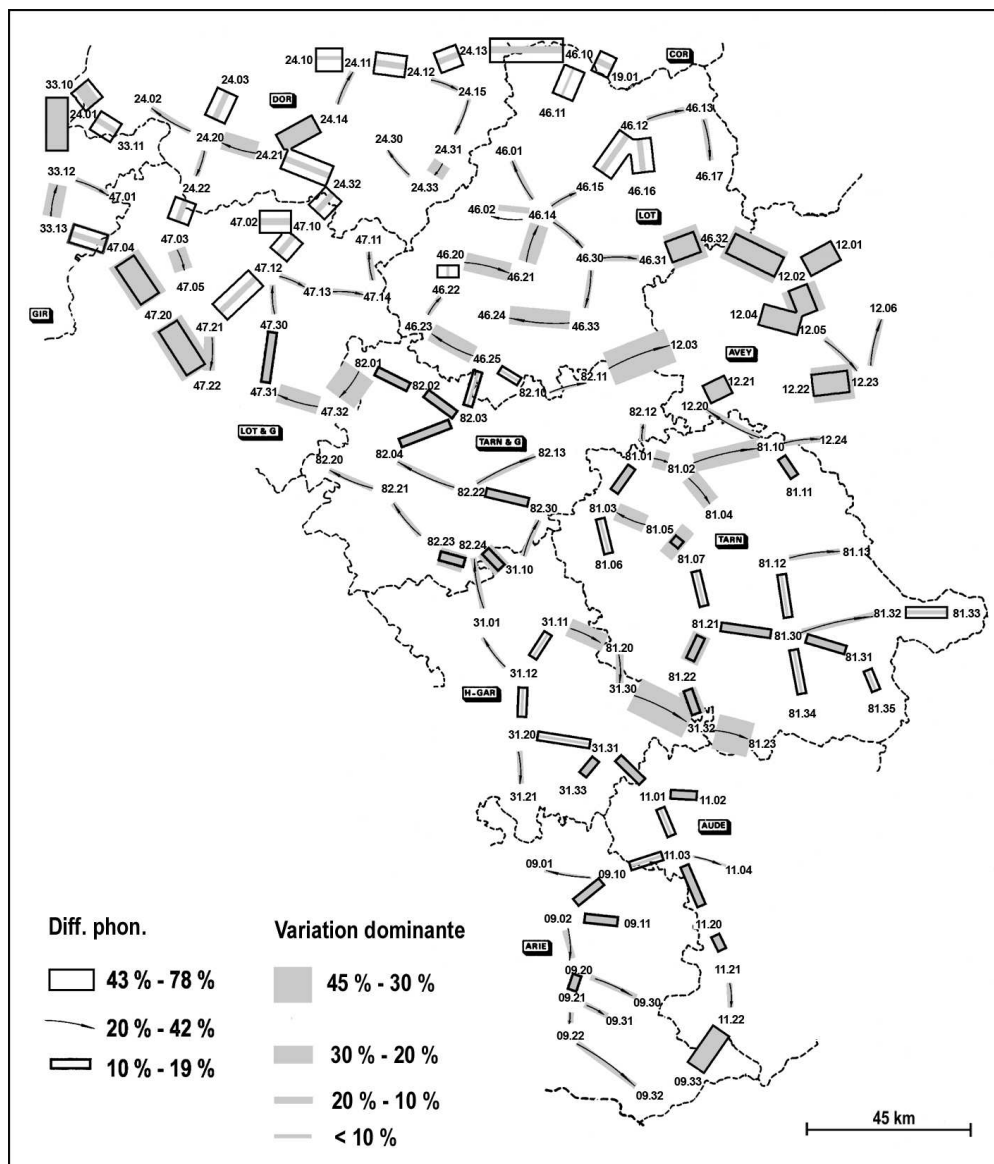


Two possible explanations can account for a large phonetic difference (col F.): First, most cases in the northern area result from an accumulation of variations (col I, K, M, large rectangles + small gray lines, map 4), see e.g. 47.10 > 24.32 or 24.11 > 24.10 representing an extreme case of scattering. Secondly, a lesser represented difference, the preponderance of a single variation (neck. F, I, big gray rectangles, map 4), see 46.32 > 12.02 or two, see 47.22 > 47.20, is reflected on the eastern and western margins of the area and the segment neighbouring the Catalan area (11.22 > 09.33).

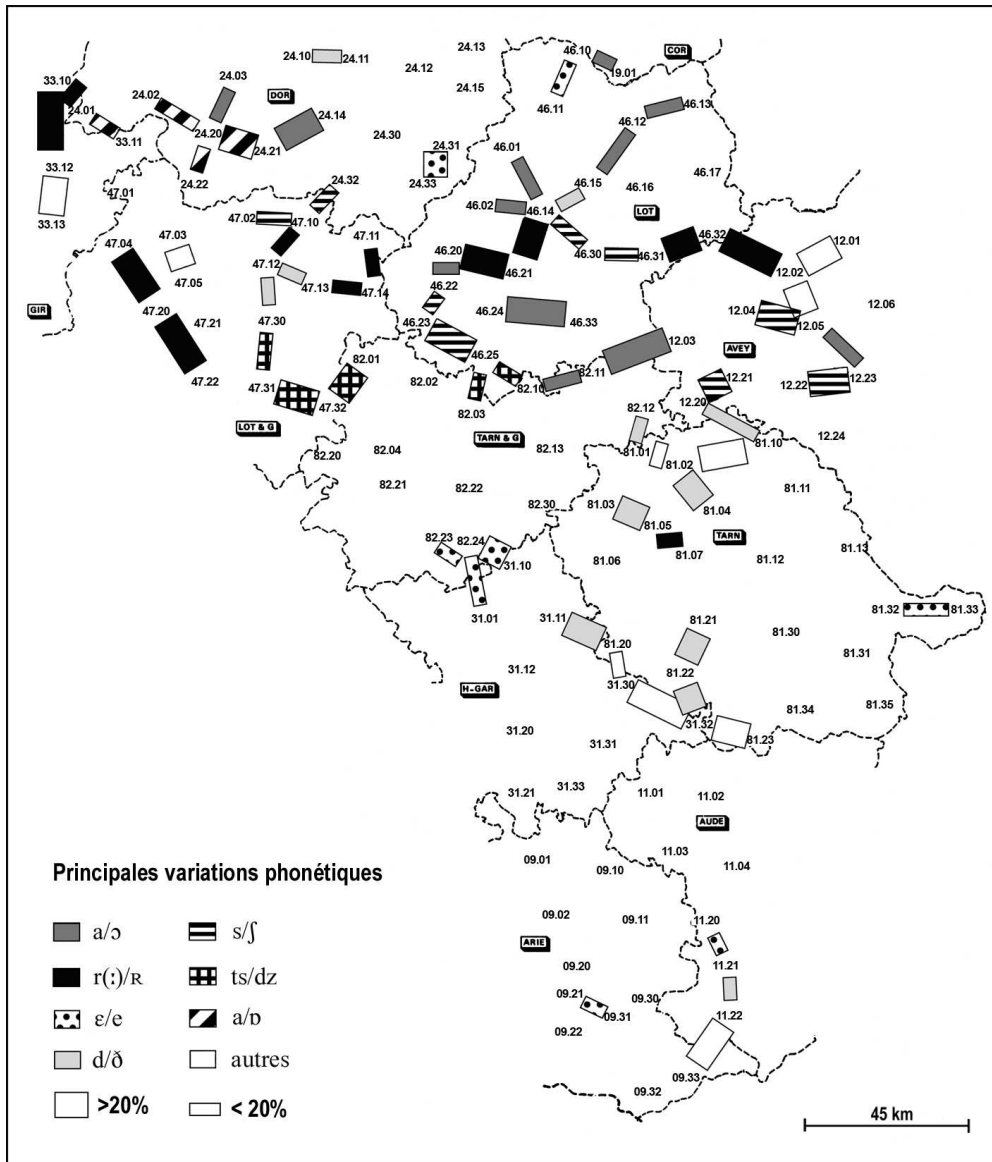
In the segments with small phonetic differences, we notice no remarkable value for the first operation (col F, I, Small gray rectangles, Map 4), see 31.12 > 31.01, but rather a dispersion.

In contrast, in the segments with average values, particularly in the south of Lot and Lot-et-Garonne as well as in Tarn, we notice the presence of some notable variations (col I; Black arrows in gray areas, map 4) see 31.32 > 81.23 or 82.11 > 12.03.

Finally, the phonemes plotted on map 5 show a strong presence of alternating r (:)/ □ and a/□ in the northern area, as foreshadowed in table 2, while eastern Tarn is characterized by the presence of d/δ. Note also that some points are the centre of a large number of variations such as 46.14 or 24.20 which stand out from their neighbours.



Map 4: phonetic differences and proportion of the main variation



Map 5: Key phonetic variations

## 5 Conclusions

Inputting unpublished Alloc lexical data has allowed me to test a personal version of the Damerau-Levenshtein algorithm on a limited number of pairs of locations as they were determined by the initial choice of the corpus. Beyond the accumulation of numbers usually generated by statistic measurements it seemed interesting to materialize the obstacles to intercomprehension between the various segments studied.

It would be instructive to extend this research to all the segments in a full triangulation, which would reveal further divisions inside the studied area<sup>17</sup>.

<sup>17</sup> I also plan in the near future, with the support of Hans Goebel, to test the data using Salzburg's DM system.

## Bibliography

- 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<sup>e</sup> 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 17<sup>e</sup> 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-).
- Goebel, 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.
- 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 spatiale 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.

## ANNEXE : tableau récapitulatif

A : segments (Figures in brackets indicate the average)

B : C. Camps differences (gross) in %

C: GBT differences in %

D : Difference between Camps and GBT

E : GBT lexical differences %

F : GBT phonetic differences %

G : GBT total number of operations

H : operation 1 and I : H / G

J : operation 2 and K : J / G

L : operation 3 and M : L /G

N : I + K + M

Abbreviations for H, J, L : R. = replacement ; S. = suppression ; Ins. = insertion

Light gray boxes: lowest values

Dark-gray boxes: highest values

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 d <sub>3</sub> par tʃ	10	R. de r par r:	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 d <sub>3</sub> par ʒ	7	R. de tʃ par d <sub>3</sub>	7	R. de ʎ par l	5	19
0910>0902	22	21	-1	7	15	34	R. de tʃ par d <sub>3</sub>	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 d <sub>3</sub>	8	S. de s	8	27
0920>0930	32	32	0	9	25	72	R. de ʒ par ø	13	R. de ʒ par d <sub>3</sub>	6	S. de g	6	25
0921>0922	30	35	5	13	24	55	R. de l int. par l	11	R. de ø par d	9	R. de ɔ par o	7	27
0921>0931	30	37	7	13	27	54	R. de ε par e	17	R. de l int. par l	9	R. de ø 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

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 %
1101>1102	21	21	0	6	16	43	R. de y par œ	12	R. de d <sub>3</sub> 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 λ	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 λ	9	Ins. de g	6	27
1120>1121	22	23	1	12	13	27	R. de ε par e	11	R. de l par λ	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 ε	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 ε 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 λ	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 ʒ <sup>j</sup>	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 ε par e	7	R. de ʃ par ç	6	20
2411>2412	38	61	23	9	55	142	R. de ð par ts	17	R. de ʃ par ç	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 ç par ʃ	7	R. de ε 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 <sup>ɒ</sup> 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 ç par s	7	R. de ε 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 ç par ʃ	8	39
2420>2422	29	45	16	6	42	107	R. de ɒ par a	12	R. de e par ε	6	S. de ŋ	6	24
2421>2414	42	50	8	11	44	115	R. de a par ɔ	29	R. de ç 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 ç	8	R. de ε par e	4	32

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 %
2422>4703	29	55	26	10	50	123	R. de ç par s	17	R. de ɒ par a	9	R. de e par ε	7	33
2431>2433	26	30	4	2	28	72	R. de e par ε	25	R. de ɔ par a	7	R. de ʃ par ç	7	39
2432>2421	38	55	17	12	49	122	R. de θ par ð	17	R. de ʃ par ç	7	R. de β par v	6	30
2433>2430	23	38	15	3	35	97	R. de ʃ par ç	9	R. de ε par e	8	R. de ɔ par a	6	23
3101>8224	20	24	4	12	26	61	R. de ε 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 r:	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 r:	6	18
3130>3132	24	35	11	10	28	64	R. de ʒ 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(:)	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(:) par ʀ	25	R. de u par ɔ	5	R. de b par v	4	34
3312>4701	39	41	2	12	33	79	R. de u par ɔ	10	R. de ε par e	9	R. de j par ʎ	9	28
3313>3312	37	48	11	15	38	100	S. de t	22	R. de ε 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 ε	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 r: 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

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 %
4614>4615	36	46	10	10	40	96	R. de d par ð	17	R. de <sup>h</sup> par s	10	R. de ʃ par ç	7	34
4614>4630	39	43	4	6	39	106	R. de ʃ par s	13	R. de <sup>h</sup> 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(:) par ʀ	24	R. de ɔ par a	7	R. de ʃ par s	7	38
4621>4614	32	52	20	8	48	108	R. de ʀ par r(:)	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 <sup>h</sup>	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 <sup>h</sup>	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(:) par ʀ	37	R. de ʃ par s	8	R. de ts par tʃ	6	51
4632>1202	47	80	33	10	78	235	R. de ʀ par r(:)	31	R. de tʃ par çʃ	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 <sup>h</sup>	8	46
4703>4705	36	36	0	6	31	72	R. de ð par ʒ	28	R. de <sup>m</sup> par ŋ	11	R. de e par ε	10	49
4704>3313	45	50	5	11	44	122	R. de e par œ	18	R. de ʀ par r:	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 ʀ par r:	13	43
4712>4710	38	63	25	9	59	151	R. de r: par ʀ	15	R. de ε par e	10	S. de t	9	34
4712>4713	26	42	16	11	35	82	R. de ð par d	12	R. de ε par e	5	Ins. de s	4	21
4712>4721	36	58	22	13	52	116	R. de ts par ʒ	16	R. de r: par ʀ	9	R. de <sup>h</sup> par s	5	30
4713>4714	28	36	8	12	27	58	R. de r par ʀ	14	R. de ε par e	12	S. de n	10	36
4714>4711	27	39	12	7	35	84	R. de r: par ʀ	18	S. de s	13	R. de s par <sup>h</sup>	11	42
4720>4704	33	65	32	10	61	166	R. de ʀ 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(:) par ʀ	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 <sup>t</sup>	10	R. de s par <sup>s</sup>	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

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 %
4732>4731	27	35	8	11	27	57	R. de dz par ts	28	R. de h par <sup>h</sup>	11	R. de r par r:	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 <sup>h</sup>	26	Ins. de e	4	R. de l par r	4	34
8103>8101	15	19	4	7	13	28	S. de t	11	R. de g par γ	7	R. de r par r:	7	25
8103>8106	23	23	0	7	17	37	Ins. de γ	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 ʀ par r(:)	23	R. de g par γ	6	54
8107>8105	18	19	1	3	16	40	R. de r: par ʀ	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 γ par g	6	R. de <sup>h</sup> par s	5	27
8110>1224	38	41	3	9	35	90	R. de <sup>h</sup> 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 <sup>h</sup> par s	14	R. de r par l	6	R. de r: par r	6	26
8112>8113	28	48	20	12	41	95	R. de s par <sup>h</sup>	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 ʒ	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 r:	7	R. de s par h	7	S. de :	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 γ	6	R. de i par e	6	21
8131>8135	23	20	-3	11	10	23	S. de s	9	Ins. de γ	4	R. de d par ð	4	17
8132>8133	25	27	2	10	18	49	R. de e par ε	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 <sup>h</sup> par h	7	R. de s par <sup>s</sup>	6	50
8202>8201	27	25	-2	9	18	44	S. de s	11	R. de <sup>s</sup> par s	9	R. de e par ε	5	25
8203>4625	21	25	4	8	18	40	R. de ts par dz	8	R. de <sup>h</sup> par h	10	S. de k	8	26
8203>8202	20	26	6	9	19	47	S. de nasalis.	15	R. de s par <sup>s</sup>	9	Ins. de g	6	30
8204>8203	20	30	10	13	19	45	R. de s par <sup>h</sup>	16	Ins. de nasalis.	11	R. de r: par r	9	36



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 %
8210>8211	25	31	6	6	27	73	R. de a par ɔ	8	R. de a par ɔ	8	R. de <sup>h</sup> par <sup>s</sup>	8	24
8211>1203	42	46	4	11	39	98	R. de a par ɔ	32	R. de <sup>s</sup> par <sup>h</sup>	7	R. de s par <sup>s</sup>	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 ε	22	Ins. de g	5	S. de s	5	32
8224>8223	11	25	14	6	19	47	R. de e par ε	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